

(19)日本国特許庁 (J P)

(12) 公 開 特 許 公 報 (A)

(11)特許出願公開番号

特開平10-133828

(43)公開日 平成10年(1998) 5月22日

(51)Int.Cl. <sup>4</sup>	識別記号	F I
G 0 6 F 3/06	5 4 0	G 0 6 F 3/06
	3 0 5	5 4 0
		3 0 5 C

審査請求 未請求 請求項の数10 O L (全 31 頁)

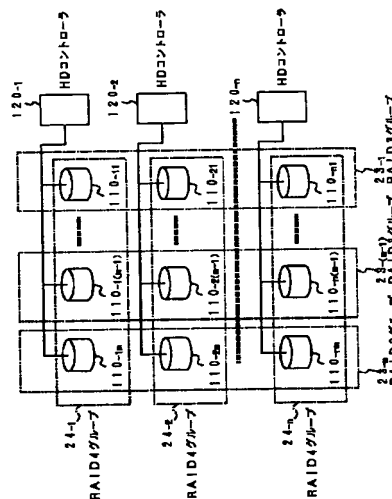
(21)出願番号	特願平8-288335	(71)出願人	000003078 株式会社東芝 神奈川県川崎市幸区堀川町72番地
(22)出願日	平成8年(1996)10月30日	(72)発明者	内堀 郁夫 東京都府中市東芝町1番地 株式会社東芝 府中工場内
		(72)発明者	金井 達徳 神奈川県川崎市幸区小向東芝町1番地 株 式会社東芝研究開発センター内
		(72)発明者	島内 芳郎 東京都府中市東芝町1番地 株式会社東芝 府中工場内
		(74)代理人	弁理士 鈴江 武彦 (外6名)

(54)【発明の名称】 マルチメディアサーバ用ディスクアレイ装置

#### (57)【要約】

【課題】同一ディスク装置上で複数のディスクアレイ方式を可能とすると共に、2重のディスクアレイ保護を可能とし、信頼性の向上を図る。

【解決手段】HDコントローラ120-1, ..., 120-nに接続される横方向配列HDD110-11 ~ 110-lm, ..., 110-n1 ~ 110-nmによりRAID4グループ24-1, ..., 24-nを、縦方向の配列HDD110-11 ~ 110-n1, ..., 110-lm ~ 110-nmによりRAID3グループ23-1, ..., 23-mをそれぞれ構成し、各HDD110-11 ~ 110-nmの領域を、長期間繰り返し利用される大容量データの格納用のRAID3&4領域と、一定期間のみ利用される大容量データの格納用のRAID3領域と、小容量データの格納用のRAID4領域とに分割し、RAID3&4領域をRAID3及びRAID4で、RAID3領域をRAID3で、RAID4領域をRAID4でそれぞれ保護する。



【特許請求の範囲】

【請求項1】 第1の方向及び第2の方向で表される2次元状に論理的に配置されたディスク装置の群であって、前記各ディスク装置のディスク領域が、第1の大容量データを格納するための第1の領域、前記第1の大容量データとは用途が異なる第2の大容量データを格納するための第2の領域、及び小容量データを格納するための第3の領域に分割して使用され、前記第1の方向のディスク装置の配列毎に、その配列内の全ての前記ディスク装置が並列にアクセスされる第1のディスクアレイグループが構成されると共に、前記第2の方向のディスク装置の配列毎に、その配列内の選択された前記ディスク装置がアクセスされる第2のディスクアレイグループが構成されるディスク装置の群と、前記第2の方向（横方向）の前記ディスク装置の配列毎に設けられ、対応する配列内の前記各ディスク装置をアクセス制御するディスクコントローラと、ホスト装置からの要求に従い前記各ディスクコントローラを制御する制御手段と、前記制御手段の制御により前記各ディスクコントローラとの間のデータ入出力を行う入出力手段であって、前記各ディスクコントローラから読み出されたデータに基づくエラー修正と、前記ホスト装置から与えられる書き込みデータのエラー修正情報である第1のタイプのパリティの生成が可能な入出力手段とを具備し、前記第1のディスクアレイグループ内の前記各ディスク装置の前記第1の領域には前記第1の大容量データの分割データまたはそのエラー修正情報である第1のタイプのパリティが配置され、前記第1のディスクアレイグループ内の前記各ディスク装置の前記第2の領域には前記第2の大容量データの分割データまたはそのエラー修正情報である第1のタイプのパリティが配置され、前記第2のディスクアレイグループ内の前記各ディスク装置の前記第3の領域には前記小容量データの分割データまたはそのエラー修正情報である第2のタイプのパリティが配置され、前記第2のディスクアレイグループ内の前記ディスク装置の前記第1の領域には、同一グループ内の他の前記各ディスク装置の前記第1の領域に配置された分割データのエラー修正情報である第2のタイプのパリティが配置されることを特徴とするマルチメディアサーバ用ディスクアレイ装置。

【請求項2】 前記制御手段は、前記ホスト装置により前記第1の領域からのデータ読み出しが要求された場合、前記各ディスクコントローラに対して、目的とする前記第1のディスクアレイグループ内の前記各ディスク装置の前記第1の領域からのデータ読み出しを行わせると共に、読み出しエラー時には、該当するディスク装置が属する前記第2のディスクアレイグループ内の他の前

記各ディスク装置の対応する前記第1の領域からデータを読み出してそのデータに基づくエラー修正を行わせる第1の方式によるリカバリ有りモードの読み出し制御を行い、エラー修正不可のディスクコントローラが1台だけ存在する場合には、他の前記各ディスクコントローラから読み出されたデータに基づくエラー修正を前記入出力手段に行わせるように構成されていることを特徴とする請求項1記載のマルチメディアサーバ用ディスクアレイ装置。

【請求項3】 前記制御手段は、前記ホスト装置により前記第1の領域からのデータ読み出しが要求された場合で且つ品質よりコンスタンス性を重視するとき、または前記ホスト装置により前記第2の領域からのデータ読み出しが要求された場合は、前記各ディスクコントローラに対して、目的とする前記第1のディスクアレイグループ内の前記各ディスク装置の前記第1の領域または前記第2の領域からのデータ読み出しのみを行わせ、読み出しエラーとなったディスクコントローラが1台だけ存在する場合には、他の前記各ディスクコントローラから読み出されたデータに基づくエラー修正を前記入出力手段に行わせるように構成されていることを特徴とする請求項2記載のマルチメディアサーバ用ディスクアレイ装置。

【請求項4】 前記制御手段は、前記ホスト装置により前記第1の領域へのデータ書き込みが要求された場合、前記ホスト装置から与えられる書き込みデータの前記第1のタイプのパリティを前記入出力手段により生成させると共に、前記書き込みデータの分割データまたは前記入出力手段により生成された第1のタイプのパリティを前記各ディスクコントローラにより目的とする前記第1のディスクアレイグループ内の前記各ディスク装置の前記第1の領域に書き込ませると共に、当該各ディスク装置がそれぞれ属する前記第2のディスクアレイグループ内の前記各ディスク装置の対応する前記第1の領域のデータに基づく前記第2のタイプのパリティを生成させて対応するディスク装置の第1の領域に書き込ませることを特徴とする請求項1記載のマルチメディアサーバ用ディスクアレイ装置。

【請求項5】 前記第1の領域及び前記第2の領域には、前記ディスク装置のディスク領域において、前記第3の領域より外周側の領域が割り当てられることを特徴とする請求項1記載のマルチメディアサーバ用ディスクアレイ装置。

【請求項6】 前記制御手段は、前記ホスト装置により前記第3の領域からのデータ読み出しが要求された場合、前記ホスト装置により指定されたディスク装置に対応する前記ディスクコントローラに対し、当該ディスク装置の前記第3の領域からデータを読み出すための第2の方式によるリカバリ無しモードの読み出しを行わせ、読み出しエラーが発生したときには、前記指定されたデ

ディスク装置が属する前記第2のディスクアレイグループ内の他の前記各ディスク装置の対応する前記第3の領域からデータを読み出してそのデータに基づくエラー修正を行うための第2の方式によるリカバリモードの読み出しを行わせ、読み出しエラーのためにエラー修正不可のときには、前記指定されたディスク装置が属する前記第2のディスクアレイグループ内の前記指定されたディスク装置からのデータ読み出しと、読み出しエラー時に前記指定されたディスク装置が属する前記第2のディスクアレイグループ内の他の前記各ディスク装置の対応する前記第3の領域からデータを読み出してそのデータに基づくエラー修正とを行うための第2の方式によるリカバリモードの読み出しを行わせるようにスケジュールすることを特徴とする請求項1記載のマルチメディアサーバ用ディスクアレイ装置。

【請求項7】 前記制御手段は、前記ホスト装置により前記第1の領域からのデータ読み出しが要求された場合、前記各ディスクコントローラに対して、目的とする前記第1のディスクアレイグループ内の前記各ディスク装置の前記第1の領域からのデータ読み出しのみ行わせる第2の方式によるリカバリ無しモードの読み出し制御を行い、読み出しエラーとなったディスクコントローラが1台だけ存在する場合には、他の前記各ディスクコントローラから読み出されたデータに基づくエラー修正を前記入出力手段に行わせ、読み出しエラーとなったディスクコントローラが2台以上存在する場合には、その読み出しエラーとなった各ディスクコントローラに対して、対応する前記第2のディスクアレイグループ内の該当する前記ディスク装置の前記第1の領域からのデータ読み出しを行わせると共に、読み出しエラー時には、当該第2のディスクアレイグループ内の他の前記各ディスク装置の対応する前記第1の領域からデータを読み出してそのデータに基づくエラー修正を行わせる第2の方式によるリカバリ有りモードの読み出し制御を行い、エラー修正不可のディスクコントローラが1台だけ存在する場合には、他の前記各ディスクコントローラから読み出されたデータに基づくエラー修正を前記入出力手段に行わせるように構成されていることを特徴とする請求項1記載のマルチメディアサーバ用ディスクアレイ装置。

【請求項8】 前記制御手段は、前記第1の領域への入出力要求と前記第2の領域への入出力要求との待ち行列である第1のキューを前記第1のディスクアレイグループ毎に、前記第1の領域に対する前記第2の方式によるリカバリ有りモードの読み出し要求と、前記第2のディスクアレイグループ内の前記各ディスク装置の前記第1の領域のデータに基づく前記第2のタイプのパリティを生成させて対応するディスク装置の第1の領域に書き込ませるための第2の方式によるパリティ生成・書き込み要求との待ち行列である第2のキューを前記第2のディスクアレイグループ毎に、それぞれ有し、

前記ホスト装置により前記第1の領域へのデータ書き込みが要求された場合には、前記ホスト装置から与えられる書き込みデータの前記第1のタイプのパリティを前記入出力手段により生成させると共に、前記書き込みデータの分割データまたは前記入出力手段により生成された前記第1のタイプのパリティを前記各ディスクコントローラにより目的とする前記第1のディスクアレイグループ内の前記各ディスク装置の前記第1の領域に書き込ませるための入出力要求を当該第1のディスクアレイグループに対応する前記第1のキューにつなぎ、この第1のキューにつないだ前記入出力要求の示す書き込みが正常終了したならば、前記第1のディスクアレイグループ内の前記各ディスク装置がそれぞれ属する前記各第2のディスクアレイグループ毎に、当該第2のディスクアレイグループを対象とする前記第2の方式によるパリティ生成・書き込み要求を、当該第2のディスクアレイグループに対応する前記第2のキューにつなぎ、前記ホスト装置に正常終了を通知するように構成されていることを特徴とする請求項7記載のマルチメディアサーバ用ディスクアレイ装置。

【請求項9】 前記制御手段は、前記第2の方式によるパリティ生成・書き込み要求を、同一の前記第2のキューの中で、どの前記第2の方式によるリカバリ有りモードの読み出し要求よりも先に実行される側につなぐことを特徴とする請求項8記載のマルチメディアサーバ用ディスクアレイ装置。

【請求項10】 前記制御手段は、前記第3の領域への入出力要求の待ち行列である第3のキューを前記第2のディスクアレイグループ毎に更に有すると共に、前記第1のキューにつなされる要求の実行に要する時間に余裕値を加えた予め定められた第1の時間を記憶しておくための第1の時間記憶手段と、前記第1の時間、及び前記第2のキュー並びに前記第3のキューにつなされる各要求の実行に要する時間の上限値より大きい第2の時間を記憶しておくための第2の時間記憶手段と、前記第1の時間と実際の要求実行に要する時間の差の累積値である第3の時間を記憶しておくための第3の時間記憶手段とを更に有しており、

前記第1のキューにつなごうとする前記第1の領域または前記第2の領域への入出力要求の示す転送長が予め定められた一定長以上の場合には、当該入出力要求を前記一定長以下の転送長の複数の入出力要求に分割して前記第1のキューにつなぎ、前記第2のキュー及び前記第3のキューにつなぐ要求中には、その要求を他の要求との干渉がない状態で実行するのに要する時間の上限値を設定し、

通常は、前記第1のキューにつながれている要求を最優先として前記第1のディスクアレイグループが重複しないようにスケジュールすると共に、前記第2のキュー及び前記第3のキューにつながれている要求については、

その要求中に設定されている前記上限値が前記第1の時間より小さいものを対象にスケジュールし、少なくとも1つスケジュールできたならば、前記第1の時間を前記第3の時間に加えた値を新たな第3の時間として前記第3の時間記憶手段に記憶すると共に、そのスケジュールした要求の実行を制御し、その実行が全て終了するまでの経過時間を前記第3の時間から差し引いた値を新たな前記第3の時間として前記第3の時間記憶手段に記憶し、1つもスケジュールできなかったならば、前記第2の時間を新たな前記第3の時間として前記第3の時間記憶手段に記憶し、前記第3の時間が前記第2の時間以上となった場合には、前記第2のキュー及び前記第3のキューにつながれている要求のみを対象にスケジュールし、少なくとも1つスケジュールできたならば、そのスケジュールした要求の実行を制御して、その実行が全て終了した後に、1つもスケジュールできなかったならば直ちに、前記第3の時間記憶手段に初期値0を記憶するように構成されていることを特徴とする請求項9記載のマルチメディアサーバ用ディスクアレイ装置。

#### 【発明の詳細な説明】

##### 【0001】

【発明の属する技術分野】本発明は、複数のディスクドライブ（ディスク装置）を備えたマルチメディアサーバ用ディスクアレイ装置に関する。

##### 【0002】

【従来の技術】近年、複数のディスクドライブを備え、これら各ディスクドライブ（に装着されている記憶媒体）にデータを分散して格納することで、並列アクセスを可能としてアクセスの高速化を図ったディスクアレイ装置が開発されている。この種のディスクアレイ装置の代表的なものとしてRAID（Redundant Arrays of Inexpensive Disk）と称されるアーキテクチャを適用したディスクアレイ装置、即ちRAID装置が知られている。このRAIDアーキテクチャは、次のような文献、D. Patterson, G. Gibson, and R. Katz, "A Case for Redundant Arrays of Inexpensive Disk (RAID)", ACM SIGMOD conference proceedings, Chicago, IL., June 1-3, 1988、及び「Randy H. Katz, Garth A. Gibson, and David A. Patterson, "Disk System Architectures for High Performance Computing," Report No. UCB/CSD 89/497 March 1989」に記載されている。

【0003】RAIDアーキテクチャは、上記したようにデータを複数のディスクドライブに分散して格納することでアクセスの高速化を図る他、その分散格納されるデータに対応してパリティと称されるエラー修正情報（冗長データ）を格納することで、いずれかのディスクドライブに障害が発生したときに、パリティと残りの正常なディスクドライブのデータとから障害ディスクドライブのデータを回復（リカバリ）して信頼性の向上を図るようにしたものである。

【0004】ところで、上記したようなディスクアレイ装置の利用形態の1つに、マルチメディア情報を格納する装置（マルチメディアサーバ）としての使用がある。このマルチメディア情報を格納する装置には、次のような3つの要求がある。

##### （1）信頼性の向上

まず、ビデオデータのように大量のデータを格納するためには、従来から知られているRAID装置（ディスクアレイ装置）以上に、冗長度が小さく、且つ信頼性が高い装置が要求されている。

【0005】例えば、ディスクドライブの障害が発生した場合、RAIDで保護されていない場合には、元のデータを再投入する必要がある。ところがビデオデータなどは、大量データであるため、ストライピング（データストリームを一定の単位で分割して分散配置すること）が行われることが多く、障害が発生したディスクドライブのデータだけでなく、関係する全てのデータの再投入を行う必要がある。この場合には、データのリカバリのために膨大な時間が必要となる。

【0006】そのため、RAIDの保護範囲を広げて一層の信頼性の向上を図ることで、このような事態の発生を防止することが要求される。

##### （2）異種データの混在

RAIDアーキテクチャ（RAID機構）は主にRAID0～5（RAIDレベル0～5）の6つ（のレベル）に分類される。この6つのレベルのうち、RAID3～5について簡単に述べる。

【0007】まずRAID3は、比較的小さい単位（例えばバイト単位）でのストライピングを行うことで、1つのI/O要求に対して全ディスクドライブを並列（同時）にアクセスできるようにすると共に、ストライピングされたデータのパリティを専用のディスクドライブ（パリティ・ディスク）に格納する方式である。

【0008】RAID4は、比較的大きい単位（例えばディスクの物理セクタ以上の単位）でのストライピングを行うことで、複数のI/O要求の対象となるディスクドライブが要求毎に異なるようにして、各要求に対応するディスクドライブを独立にアクセスできるようにすると共に、ストライピングされたデータのパリティを専用のディスクドライブに格納する方式である。

【0009】RAID5は、ストライピングされたデータとパリティとがインターリーブされる点を除いてRAID4と同様である。以上のことから、ビデオデータのように、大量で、しかもディスクドライブの障害時にも一定の転送レートを保証しなければならないデータの格納には、RAID3が適している。

【0010】一方、静止画のように、比較的小量で、スループットが重視されるデータの格納には、RAID4または5が適している。ところが、ビデオデータと静止画の両データを扱う場合には、RAID3またはRAI

D4 (5) の一方だけでは不具合が生じる。

【0011】このため、同一RAID装置（ディスクアレイ装置）上で、RAID3とRAID4 (5) の両機構（レベル）を実現することが要望される。

(3) ビデオデータに要求される、多様な、保全性と要求品質への対応

ビデオデータの利用形態にも、例えば、(a) 監視システムなどでの応用のように一定時間のみの利用され、必要ならバックアップが取られた後、上書きされるようなデータと、(b) 長時間保存され、繰り返し利用されるデータとがあり、それぞれのデータで、耐障害性への要求が異なる。つまり、付加する冗長度、及び許されるオーバヘッドが異なる。

【0012】また、ビデオデータの送出（転送）形態にも、例えば、(a) バッファメモリの少ないユーザ端末へ直接送出されるケースと、(b) 他のサーバ（ディスクアレイ装置）へ送られ（た後、そこからユーザ端末へ繰り返し送出され）るケースとがあり、それぞれのケースで、送出のための読み出し時の要求品質、及びコンスタント性が異なる。

【0013】このため、同一RAID装置（ディスクアレイ装置）上で、上記の多様な要求に応えることが要望される。そこで従来は、図20或いは図21に示するようなディスクアレイ装置が考えられている。

【0014】図20のディスクアレイ装置は、複数のディスクドライブ（ディスク装置）、例えば複数のハードディスクドライブ（HDD）201を論理的に2次元状に配置して、コントローラ202の制御のもとで2次元のRAID方式を実現したものである。

【0015】この図20のディスクアレイ装置では、2重にRAID保護が図られることから、上記(1)の要求には応えられるが、上記(2)、(3)の要求には応えられない。

【0016】一方、図21のディスクアレイ装置は、RAID3で制御されるHDD211の列と、RAID4、5で制御されるHDD212の列とを有し、RAID3とRAID4、5の双方の機能を持つコントローラ213により、これらHDD211の列（RAID3のグループ）とHDD212の列（RAID4、5のグループ）を制御する構成としたものである。

【0017】この図22のディスクアレイ装置では、信頼性の点で改善されておらず、上記(1)の要求に応えていない。また、ディスクの負荷が均一にならず効率が悪く、ビデオデータなどの大量のデータと静止画などの小量のデータとの割合が変化するシステムではディスク内のスペースの使用効率が悪化する。つまり、上記

(2)の要求に応えられない。また、上記(3)のような柔軟性（多様性）への要求にも応えられない。

【0018】

【発明が解決しようとする課題】このように従来のディ

スクアレイ装置では、上記(1)、(2)、(3)で示したような要求に応えることができなかった。この発明は上記事情を考慮してなされたものでその目的は、ディスク装置の論理的な配列の方向によって異なる機能を持つ2次元のディスクアレイを実装することで、同一ディスク装置上で複数のディスクアレイ方式を可能とすると共に、2重のディスクアレイ保護を可能とし、信頼性の向上を図ると共に、異種データの混在を可能とし、しかも各データ種類・用途に応じた最適なアクセスが行えるマルチメディアサーバ用ディスクアレイ装置を提供することにある。

【0019】

【課題を解決するための手段】本発明のディスクアレイ装置は、第1の方向及び第2の方向で表される2次元状に論理的に配置されたディスク装置の群であって、各ディスク装置のディスク領域が、第1の大容量データを格納するための第1の領域（以下、RAID3&4領域と称する）、上記第1の大容量データとは用途が異なる第2の大容量データ（以下、RAID3領域と称する）を格納するための第2の領域、及び小容量データを格納するための第3の領域（以下、RAID4領域と称する）に分割して使用され、上記第1の方向のディスク装置の配列（以下、第1の方向の配列と称する）毎に、その配列内の全てのディスク装置が並列にアクセスされる第1のディスクアレイグループ（以下、RAID3グループと称する）が構成されると共に、上記第2の方向のディスク装置の配列（以下、第2の方向の配列と称する）毎に、その配列内の選択されたディスク装置がアクセスされる第2のディスクアレイグループ（RAID4グループ）が構成されるディスク装置の群と、上記第2の方向の配列毎に設けられ、対応する配列内の各ディスク装置をアクセス制御するディスクコントローラと、ホスト装置からの要求に従い各ディスクコントローラを制御する制御手段と、この制御手段の制御により各ディスクコントローラとの間のデータ入出力を行う入出力手段（RAID機構）であって、各ディスクコントローラから読み出されたデータに基づくエラー修正（以下、RAID3によるリカバリと称する）と、ホスト装置から与えられる書き込みデータのエラー修正情報である第1のタイプのパリティ（以下、RAID3のパリティと称する）が生成可能な入出力手段（RAID機構）とを備えており、上記RAID3グループ（第1のディスクアレイグループ）内の各ディスク装置のRAID3&4領域には上記第1の大容量データの分割データまたはそのエラー修正情報であるRAID3のパリティ（第1のタイプのパリティ）が配置され、上記第1のディスクアレイグループ内の各ディスク装置のRAID3領域には上記第2の大容量データの分割データまたはそのエラー修正情報であるRAID3のパリティ（第1のタイプのパリティ）が配置され、上記RAID4グループ（第2のディ

スクアレグループ) 内の各ディスク装置のRAID4領域には上記小容量データの分割データまたはそのエラー修正情報である第2のタイプのパリティ(以下、RAID4のパリティと称する)が配置され、上記RAID4グループ内のディスク装置のRAID3&4領域には、同一グループ内の他の各ディスク装置のRAID3&4領域に配置された分割データのエラー修正情報であるRAID4のパリティが配置されることを特徴とする。

【0020】このようなディスクアレイ装置においては、第1の大容量データが格納されるRAID3&4領域はRAID3とRAID4の両方で保護される。したがってRAID3&4領域は、耐障害性への要求がより高い、例えば長期間繰り返し利用される通常のビデオデータの格納に適している。また、第2の大容量データが格納されるRAID3領域はRAID3で保護される。したがってRAID3領域は、上記第1の大容量データに比べれば耐障害性への要求は低いが、オーバーヘッドがより少ないことが要求される、例えば監視システムでの応用のように一定期間のみ利用されるビデオデータの格納に適している。一方、スループットが重視される小容量データ、例えば静止画等のデータは、各ディスク装置のRAID4領域に格納することでRAID4が適用可能となるため、スループットの向上が図れる。

【0021】また本発明は、上記ディスクアレイ装置において、第1の大容量データの読み出し時の耐障害性を実現するため、ホスト装置により上記RAID3&4領域からのデータ読み出しが要求された場合に、制御手段は、上記各ディスクコントローラに対して、目的とするRAID3グループ内の各ディスク装置のRAID3&4領域からのデータ読み出しを行わせると共に、読み出しエラー時には、該当するディスク装置が属するRAID4グループ内の他の各ディスク装置の対応するRAID3&4領域からデータを読み出してそのデータに基づくエラー修正を行わせる第1の方式によるリカバリ有りモードの読み出し(以下、RAID4のリカバリ有りの読み出しと称する)制御を行い、エラー修正不可のディスクコントローラが1台だけ存在する場合には、他の前記各ディスクコントローラから読み出されたデータに基づくエラー修正(RAID3によるリカバリ)を上記入出力手段に行わせることを特徴とする。

【0022】このようなディスクアレイ装置においては、RAID3による読み出しでエラーとなったディスク装置が複数存在したとしても、RAID4による読み出しを行ってRAID4のパリティを用いたエラー修正を実行することでリカバリすることが可能となる。但し、RAID3による読み出しでエラーとなったディスク装置が1つだけの場合には、上記記入出力手段(RAID機構)でのRAID3によるリカバリで対処可能なため、対応するディスクコントローラ1台のみが行うリカ

バリ処理に要する時間が無駄となる。

【0023】そこで、このような無駄を無くすため、ホスト装置によりRAID3&4領域からのデータ読み出しが要求された場合に、上記制御手段は、上記各ディスクコントローラに対して、目的とするRAID3グループ内の各ディスク装置のRAID3&4領域からのデータ読み出しのみ行わせる第2の方式によるリカバリ無しモードの読み出し(RAID4によるリカバリ無しモードの読み出し)制御を行い、読み出しエラーとなったディスクコントローラが1台だけ存在する場合には、他の各ディスクコントローラから読み出されたデータに基づくエラー修正(RAID3による復旧)を入出力手段(RAID機構)に行わせ、読み出しエラーとなったディスクコントローラが2台以上存在する場合に、その読み出しエラーとなった各ディスクコントローラに対して、対応するRAID4グループ内の該当するディスク装置のRAID3&4領域からのデータ読み出しを行わせると共に、読み出しエラー時には、当該RAID4グループ内の他の各ディスク装置の対応するRAID3&4領域からデータを読み出してそのデータに基づくエラー修正を行わせる第2の方式によるリカバリ有りモードの読み出し(RAID4によるリカバリ有りモードの読み出し)制御を行い、エラー修正不可のディスクコントローラが1台だけ存在する場合には、他の各ディスクコントローラから読み出されたデータに基づくエラー修正を上記入出力手段に行わせることを特徴とする。

【0024】また本発明は、ホスト装置によりRAID3&4領域からのデータ読み出しが要求された場合でも、品質よりコンスタンス性を重視するとき、例えば読み出したデータを直接ユーザに送信するような場合には、他のサーバに送信するために読み出す場合と異なって、RAID3領域からのデータ読み出しが要求された場合と同様に、RAID3による読み出しのみを行う、即ち上記各ディスクコントローラに対して、目的とするRAID3グループ内の各ディスク装置の前記第1の領域からのデータ読み出しのみを行わせ、読み出しエラーとなったディスクコントローラが1台だけ存在する場合には、他の前記各ディスクコントローラから読み出されたデータに基づくエラー修正を上記入出力手段に行わせることを特徴とする。

【0025】このように本発明においては、RAID3とRAID4とで2重に保護されているRAID3&4領域から、用途に応じて、2種類の読み出し方法のいずれかが選択可能である。

【0026】また本発明は、ホスト装置によりRAID4領域からのデータ読み出しが要求された場合に、ホスト装置により指定されたディスク装置に対応するディスクコントローラに対し、当該ディスク装置のRAID4領域からのデータ読み出しのみ行わせ(RAID4によるリカバリ無しモードの読み出しを行わせ)、読み出し

エラーが発生したときには、上記指定されたディスク装置が属するRAID4グループ内の他の各ディスク装置の対応するRAID4領域からのデータ読み出しとそのデータに基づくエラー修正を行わせ（RAID4によるリカバリモードの読み出しを行わせ）、読み出しエラーのためにエラー修正不可のときには、上記指定されたディスク装置が属するRAID4グループ内の当該ディスク装置からのデータ読み出しと、読み出しエラー時に当該ディスク装置が属するRAID4グループ内の他の各ディスク装置の対応するRAID4領域からデータを読み出してそのデータに基づくエラー修正を行わせる

（RAID4によるリカバリ有りモードの読み出しを行わせる）ようにスケジュールすることを特徴とする。

【0027】このように、動作するディスク装置を細かく多段階に制御することで、大容量データの読み出しのスケジューリングに影響を与えないように、RAID4領域からの小容量データの読み込みを行う機会を増加させることが可能となる。

【0028】また本発明は、大容量データが格納されるRAID3&4領域及びRAID3領域を、小容量データが格納されるRAID4領域より外周側のディスク領域に割り当てたことを特徴とする。

【0029】このように、小容量データをディスクの内周部に、大容量データをそれより外周部に配置することにより、即ち内周部に比べてトラック当たりのデータ量が多い領域に大容量データを配置することにより、大容量データの配信の性能を向上させることができる。

【0030】また本発明は、ホスト装置によりRAID3&4領域へのデータ書き込みが要求された場合に、上記制御手段は、ホスト装置から与えられる書き込みデータの第1のタイプのパリティ（RAID3のパリティ）を上記入出力手段により生成させると共に、当該書き込みデータの分割データまたは上記生成されたパリティを上記各ディスクコントローラにより目的とするRAID3グループ内の各ディスク装置のRAID3&4領域に書き込ませると共に、当該各ディスク装置がそれぞれ属する各RAID4グループ内の各ディスク装置の対応するRAID3&4領域のデータに基づく第1のタイプのパリティ（RAID4のパリティ）を生成させて対応するディスク装置のRAID3&4領域に書き込ませることを特徴とする。

【0031】このように本発明においては、RAID3&4領域に書き込まれるデータに対しては、RAID3のパリティとRAID4のパリティの2種のパリティが生成され、RAID3とRAID4とで2重に保護できる。

【0032】また本発明は、上記したRAID3&4領域へのデータ書き込みでは、RAID4のパリティの生成・書き込みに時間を要することから、その不具合を解消するために、RAID3&4領域への入出力要求と

RAID3領域への入出力要求との待ち行列である第1のキュー（キュー#1）をRAID3グループ毎に、RAID3&4領域に対するRAID4によるリカバリ有りモードの読み出し要求と、RAID4によるパリティ生成・書き込み要求との待ち行列である第2のキュー

（キュー#2）をRAID4グループ毎に、それぞれ上記制御手段に持たせ、ホスト装置によりRAID3&4領域へのデータ書き込みが要求された場合には、ホスト装置から与えられる書き込みデータのRAID3のパリティを上記入出力手段により生成させると共に、当該書き込みデータの分割データまたは上記生成されたパリティを上記各ディスクコントローラにより目的とするRAID3グループ内の各ディスク装置のRAID3&4領域に書き込ませるための入出力要求を当該RAID3グループに対応する上記第1のキューにつなぎ、この第1のキューにつないだ入出力要求の示す書き込みが正常終了したならば、上記RAID3グループ内の各ディスク装置がそれぞれ属する上記各RAID4グループ毎に、当該RAID4グループ内の各ディスク装置の対応するRAID3&4領域のデータに基づくRAID4のパリティを生成させて対応するディスク装置のRAID3&4領域に書き込ませるためのRAID4によるパリティ生成・書き込み要求を、当該RAID4グループに対応する上記第2のキューにつないで上記ホスト装置に正常終了を通知し、上記RAID4によるパリティ生成・書き込み要求の実行をホスト装置への正常終了通知より後にしたことを特徴とする。

【0033】本発明においては、RAID3&4領域へのデータ書き込み時の見かけ上の性能向上が可能となる。なお、RAID4によるパリティ生成・書き込み要求を実際に実行した結果、異常終了した場合には、ホスト装置からの要求とは非同期に制御手段からホスト装置に異常終了を通知すればよい。

【0034】また、上記RAID4によるパリティ生成・書き込み要求は、同一の第2のキューの中では、どのRAID4によるリカバリ有りモードの読み出し要求よりも先に実行される側につながれるようにすると、誤ったデータを読み込むことになるのを防ぐことができる。

【0035】また本発明は、RAID4領域への入出力要求の待ち行列である第3のキュー（キュー#3）をRAID4アレイグループ毎に制御手段に更に持たせると共に、上記第1のキューにつながれる要求の実行に要する時間に余裕値を加えた予め定められた第1の時間

（T）を記憶しておくための第1の時間記憶手段（Tレジスタ）と、上記第1の時間、及び上記第2のキュー並びに第3のキューにつながれる各要求の実行に要する時間の上限値より大きい第2の時間（A）を記憶しておくための第2の時間記憶手段（Aレジスタ）と、上記第1の時間と実際の要求実行に要する時間の差の累積値である第3の時間（t）を記憶しておくための第3の時間記

憶手段（ $t$ レジスタ）とについても制御手段に持たせ、第1のキューにつなごうとするRAID3&4領域またはRAID3領域への入出力要求の示す転送長が予め定められた一定長以上の場合には、当該入出力要求を上記一定長以下の転送長の複数の入出力要求に分割して第1のキューにつなぎ、第2のキュー及び第3のキューにつなぐ要求中には、その要求を他の要求との干渉がない状態で実行するのに要する時間の上限値を設定し、通常は、第1のキューにつながれている要求を最優先としてRAID3グループが重複しないようにスケジュールすると共に、第2のキュー及び第3のキューにつながれている要求については、その要求中に設定されている上限値が上記第1の時間（ $T$ ）より小さいものを対象にスケジュールし、少なくとも1つスケジュールできたならば（ここでのスケジュールを第1のスケジュールと称する）、上記第1の時間（ $T$ ）を上記第3の時間（ $t$ ）に加えた値を新たな第3の時間（ $t$ ）として上記第3の時間記憶手段に記憶すると共に、そのスケジュールした要求の実行を制御し、その実行が全て終了するまでの経過時間を上記第3の時間（ $t$ ）から差し引いた値を新たな第3の時間（ $t$ ）として上記第3の時間記憶手段に記憶し、1つもスケジュールできなかったならば、上記第2の時間（ $A$ ）を新たな第3の時間（ $t$ ）として上記第3の時間記憶手段に記憶し、上記第3の時間（ $t$ ）が上記第2の時間（ $A$ ）以上となった場合に限り、第2のキュー及び第3のキューにつながれている要求のみを対象にスケジュールし、少なくとも1つスケジュールできたならば（ここでのスケジュールを第2のスケジュールと称する）、そのスケジュールした要求の実行を制御して、その実行が全て終了した後に、1つもスケジュールできなかったならば直ちに、上記第3の時間記憶手段に初期値0を記憶するようにしたことを特徴とする。

【0036】本発明において、上記第1のスケジュールにより対応する要求が実行されると、その要求の実行の経過時間は $T$ （第1の時間）より小さいことが保証されるため、その要求の実行後に行われる $t$ （第3の時間）の更新の結果、当該 $t$ は増加する。したがって、このような処理が繰り返される毎に、上記 $t$ は増加していき、そのうちに $t \geq A$ となる。すると、第2のキュー及び第3のキューにつながれている要求のみを対象とする第2のスケジュールが行われる。ここでは、その要求の実行に要する時間の上限値が $T$ より大きいもの（但し、 $A$ よりは小さい）もスケジュールの対象となる。この第2のスケジュールの結果、スケジューリングされた要求が実行されると、上記 $t$ は初期値0に更新されるため、今度は第1のスケジュールが行われることになり、第2のスケジュールが続けて行われることはない。

【0037】したがって本発明においては、第1のキューが空きの状態で新しい要求が登録された場合には、その要求は（ $A+T$ ）時間以内にスケジュールされる。ま

た1つの第1のキューに例えば $p$ 個の要求がつながっている状態で、新しい要求が登録された場合には、先頭の要求がスケジュールされるまでの最長時間は（ $A+T$ ）であり、 $p$ 個の要求の実行に要する時間の上限は $pT$ であることから、新たに登録された要求は（ $A+(p+1)T$ ）時間以内にスケジュールされる。即ち、第1のキューへの要求の頭出し時間が保証できる。

【0038】また、本発明においては、第1のキューが空にならないようにホスト装置から制御することより、 $T$ 時間に1回の割合で入出力要求を実行することができる。即ち、第1のキューへの入出力のレートが保証できる。但し、リカバリ有りモードの入出力要求を発行した場合、第2のキューにつなごうとされた場合には、その終了の通知は遅れる。

【0039】また、本発明においては、第1のキューに要求がつながれている状態でも、第2及び第3のキューの要求にはスケジュールされる機会が訪れる。以上により、ビデオデータのような大容量データの入出力については、一定の入出力レートを保証し、且つ余裕があれば、静止画などの小容量データの入出力も並行して実行できる。

【0040】

【発明の実施の形態】以下、本発明の実施の形態につき図面を参照して説明する。図1は本発明の一実施形態に係るマルチメディアサーバ用ディスクアレイ装置10の構成を示すブロック図である。

【0041】図1のディスクアレイ装置10は、同一装置上で例えばRAID3とRAID4の両機構を実現するものであり、ディスクドライブ（ディスク装置）群、例えばハードディスクドライブ（HDD）群11と、HD（ハードディスク）コントローラ120-1～120-nからなるHDコントローラ部12と、バッファ130-1～130-nからなるバッファ部13と、RAID機構14と、制御部15とから構成される。

【0042】HDD群11は、論理的に $m \times n$ の2次元配列をなす、 $m \times n$ 個のHDD、即ちHDD110-11～110-1m、110-21～110-2m、…、110-n1～110-nmからなる。

【0043】ここで、HDD群11における $x$ 方向（横方向）の配列であるHDD110-11～110-1m、110-21～110-2m、…、110-n1～110-nmは、HDコントローラ120-1、120-2、…、120-nに接続されている。HDD110-11～110-1m、…、110-n1～110-nm（に装着される図3に示す各ディスク30の内周部IAの領域、即ちRAID4領域313）は、図2に示すように、それぞれRAID4のグループ24-1、…、24-nを構成する。また、HDD群11における $y$ 方向（縦方向）の配列であるHDD110-11～110-n1、…、110-1m～110-nm（に装着される図3に示す各ディスク30の外周部OA



の領域、即ちRAID3&4領域311とRAID3領域312)は、図2に示すように、それぞれRAID3のグループ23-1, ..., 23-mを構成する。

【0044】なお、本実施形態では、1つのHDコントローラ120-i (i=1~n)下に1つRAID4グループ24-iが接続された構成を適用しているが、全てのHDコントローラ120-i (i=1~n)に同数の複数のグループが接続される構成であっても構わない。

【0045】RAID3グループ23-mを構成するHDD110-lm ~ 110-nm (に装着される図3に示す各ディスク30の内周部IAの領域、即ちRAID4領域313と、外周部OAの領域の一部をなすRAID3&4領域311)は、RAID4グループ24-1~24-nのパリティ格納用に用いられ、RAID4グループ24-nを構成するHDD110-nl ~ 110-nm (に装着される図3に示す各ディスク30の外周部OAの領域、即ちRAID3&4領域311とRAID3領域312)は、RAID3グループ23-1~23-mのパリティ格納用に用いられる。

【0046】HDコントローラ部12内のHDコントローラ120-1~120-nは、制御部15からの指示により、RAID4のコントローラとして、自身に接続されているRAID4グループ24-1~24-n内のHDDを制御する。またHDコントローラ120-1~120-nは、制御部15からの指示により、RAID機能を持たないコントローラとしても動作する。

【0047】バッファ13内のバッファ130-1~バッファ130-nは、HDコントローラ120-1~120-nにより入出力されるデータを一時的に格納するのに用いられる。

【0048】RAID機構14は、制御部15からの指示により、RAID3の機能を実行することも、バッファ13内のバッファ130-1~130-nの内容をそのまま入出力することも可能なように構成されている。

【0049】制御部15は、図示せぬホスト装置からの指示を受け、HDコントローラ部12及びRAID機構14を通して装置全体を制御する。この制御部15は、後述する境界情報等を記憶しておくための不揮発性メモリ150を内蔵している。

【0050】さて本実施形態では、HDD群11内の各HDD110-ij (i=1~n, j=1~m)の領域、更に具体的に述べるならばHDD110-ijに装着される図3に示すようなディスク30の領域を、次の3種類に分けて使用している。

(a) RAID3及び4の双方で保護される領域311  
この領域(以下、RAID3&4領域と称する)311は、通常のビデオデータを格納するのに用いられる。この領域311への書き込みの際には、RAID3及び4の双方のパリティが生成される。

【0051】この領域311からの読み込みには、次の

2つのモードがある。第1は、直接ユーザにビデオデータを送信するために読み出すような場合に、品質よりコスタント性、システムのオーバーヘッド軽減を目指して、RAID4によるリカバリを行わないようにしたモード(以下、リカバリ無しモードと称する)である。

【0052】第2は、他のサーバ(ディスクアレイ装置)にビデオデータを送信するために読み出すような場合に、品質を重視してRAID4によるリカバリを行うようにしたモード(以下、リカバリ有りモードと称する)である。

【0053】いずれの場合にも、RAID3によるリカバリは行われる。

(b) RAID3のみで保護される領域312  
この領域(以下、RAID3領域と称する)312は、監視システムへの応用のような、短時間保存されるビデオデータなどを格納するのに用いられる。この領域312への書き込みの際には、RAID3のパリティのみが生成される。また、この領域312からの読み出しの際には、RAID3によるリカバリが行われる。

(c) RAID4のみで保護される領域313  
この領域(以下、RAID4領域と称する)313は、静止画に代表される小量のデータを格納するのに用いられる。この領域313への書き込みの際には、RAID4のパリティのみが生成される。また、この領域313からの読み出しの際には、RAID4によるリカバリが行われる。

【0054】本実施形態において、(各HDD110-ij内の)ディスク30は、内周部IAと、その外側の領域(内周部IAを除く領域)である外周部OAとに分けて管理される。内周部IAはRAID4領域313に割り当てられ、静止画等の小量データを配置(格納)するのに用いられる。したがって、内周部IAはRAID4で制御される。一方、外周部OAはRAID3&4領域311とRAID3領域312とに割り当てられ、ビデオデータ等の大量データを配置するのに用いられる。したがって、外周部OAはRAID3或いはRAID3&4で制御される。

【0055】一般に、ディスク30では、内周側より外周側の方がトラック当たりのデータ量が多いことから、外周側の方が読み出し速度(転送速度)が高速である。したがって、上記したように外周部OA側にビデオデータ等の大量データを配置する構成とすることで、ビデオデータ等の配信の性能を向上することができる。一方、静止画等の小量データは、上記したように内周部IA側に配置する構成とする。小量データは、(大量データに比べて)I/O時間(データの入出力に要する時間)の中に占めるシーク、回転待ち時間の割合が大きいので、転送時間が長くなっても性能低下の割合が少ない。

【0056】以上により、システム全体の性能が向上する。なお、図3では、RAID3領域312がRAID

3&4領域311の外側となっているが、その逆であっても構わない。

【0057】内周部IAと外周部OAとの境界の情報は、全てのHDD110-ij (i=1~n, j=1~m)内の全てのディスク30に共通であり、図1中の制御部15の有する不揮発性メモリ150内に確保された境界情報領域(以下、Bレジスタと称する)151に記憶される。この不揮発性メモリ150内のBレジスタ151の境界情報は、ホスト装置からの命令で書き換えることが可能である。

【0058】次に、本実施形態における動作を、(1)RAID3&4領域311からのリカバリ有りモードでの読み出し、(2)RAID3&4領域311への書き込み、(3)RAID3領域312からの読み出し、またはRAID3&4領域311からのリカバリ無しモードでの読み出し、(4)RAID3領域312への書き込み、(5)RAID4領域313からの読み出し、(6)RAID4領域313への書き込みについて、順に説明する。

(1)RAID3&4領域311からのリカバリ有りモードでの読み出し  
まず、RAID3&4領域311からのリカバリ有りモードでの読み出しについて、図4のフローチャートを参照して説明する。

【0059】制御部15は、ホスト装置からの指示が、通常のビデオデータ等、長期間保存されて繰り返し利用される大量データを他のサーバ(ディスクアレイ装置)に送信するためにRAID3&4領域311からの読み出しを行うものである場合、HDコントローラ部12内の各HDコントローラ120-1~120-nに対して、RAID4によるリカバリ機能有りのリードのための命令(read命令)を発行する(ステップS1)。ここで、各HDコントローラ120-i (i=1~n)に対する命令(read命令)には、リードの対象となるHDD110-ij (jは1~mのいずれか)と当該HDD110-ij内のディスク領域の情報(領域先頭位置へのシークアドレスと、転送長の情報)が含まれている。

【0060】各HDコントローラ120-iは、制御部15から与えられる命令(read命令)で要求されたHDD110-ijからのデータ読み出しを行い(ステップS2)、正常に読めたか否かをチェックする(ステップS3)。

【0061】もし、エラーがなかったならば、HDコントローラ120-iはステップS2で読み出したデータをバッファ部13内のバッファ130-iに格納し、制御部15に正常終了を通知する(ステップS4)。

【0062】もし、エラーがあったならば、HDコントローラ120-iは、RAID4によるリカバリを実行する(ステップS5)。即ちHDコントローラ120-iは、RAID4グループ24-i内のHDD110-ijを

除く他のHDD110-il~110-imを対象に、ステップS1で行ったのと同じ読み出しを行い、全て正常に読めたならば、リカバリ可能であるものと判断し、その読み出したデータの排他的論理和をとってHDD110-ijのデータをリカバリする。

【0063】HDコントローラ120-iは、RAID4によるリカバリができた場合(ステップS6)、リカバリしたデータをバッファ部13内のバッファ130-iに格納し、制御部15に正常終了を通知する(ステップS7)。これに対し、1つでも正常に読めなかったHDDが存在するならば、HDコントローラ120-iはRAID4によるリカバリはできないものとして、制御部15に対してリードエラー(リカバリ不可)を通知する(ステップS8)。

【0064】制御部15は、各HDコントローラ120-iからのエラー通知(リカバリ不可通知)を監視しており(ステップS9)、エラーを通知したHDコントローラ120-iの数が0であるならば、バッファ部13内のバッファ130-1~130-nには、HDコントローラ120-1~120-nに要求したデータが格納されているものと判断してステップS10に進む。このステップS10では、制御部15はRAID機構14に対してRAID3の機能を起動させ、バッファ130-1~130-(n-1)のデータを合成させた後、その合成後のデータをホスト装置に出力する。

【0065】また、エラーを通知したHDコントローラ120-iの数が1であるならば、制御部15はステップS11に進む。このステップS11では、制御部15はRAID機構14に対してRAID3の機能を起動させ、エラーを通知した唯一のHDコントローラ120-iを除くHDコントローラ120-1~120-nに対応するバッファ130-1~130-nのデータから、エラーを通知したHDコントローラ120-iが読み出すべきデータをバッファ130-i内にリカバリさせ、バッファ130-1~130-(n-1)のデータを合成させた後、その合成後のデータをホスト装置に出力する。

【0066】また、エラーを通知したHDコントローラ120-iの数が2以上であるならば、制御部15はRAID機構14のRAID3機能を起動してもリカバリ不可能であるとして、その旨をホスト装置に通知してエラー終了する(ステップS12)。

(2)RAID3&4領域311への書き込み  
次に、RAID3&4領域311への書き込みについて、図5のフローチャートを参照して説明する。

【0067】制御部15は、ホスト装置からの指示が、通常のビデオデータ等、長期間保存されて繰り返し利用される大量データのRAID3&4領域311への書き込みを行うものである場合、まずRAID機構14に対してRAID3の機能を起動させる。これによりRAID機構14は、ホスト装置からのデータ(書き込みデー

タ)をRAID3でストライピングして、そのパリティを生成しながら、そのストライピングしたデータ(分割データ)及びパリティを、バッファ部13内のそれぞれ対応するバッファ130-1~130-nに格納する(ステップS21)。

【0068】次に制御部15は、HDコントローラ部12内の各HDコントローラ120-1~120-nに対し、バッファ130-1~130-nに格納されたデータを、RAID4でのパリティを生成しながら、RAID3グループ23-j(jは1~m-1のいずれか)内の目的とするHDD110-1j~110-njに書き込むことを指示する(ステップS22)。

【0069】これを受けて各HDコントローラ120-i(i=1~n)は、RAID4でのパリティを生成しながら、バッファ130-iのデータをRAID3グループ23-j内の目的とするHDD110-ij(即ちRAID4グループ24-i内の目的とするHDD110-ij)に書き込むと共に、生成したパリティをRAID4グループ24-i内のHDD110-imに書き込む(ステップS23)。ここで、RAID4でのパリティの生成は次のように行われる。

【0070】まずHDコントローラ120-iは、RAID4グループ24-i内のHDD110-il~110-i(m-1)を対象に、目的HDD110-ijの書き込み対象領域に一致する領域からのデータ読み出しを行う。次にHDコントローラ120-iは、目的HDD110-ijから読み出したデータをバッファ130-iのデータで更新した後、この更新後のデータを含むHDD110-il~110-i(m-1)からの読み出しデータの排他的論理和をとってパリティを生成する。そしてHDコントローラ120-iは、上記更新後のデータをRAID4グループ24-i内の目的HDD110-ijに書き込むと共に、生成したパリティをRAID4グループ24-iのHDD110-im内の領域(目的HDDの書き込み対象領域に一致する領域)に書き込む。

(3) RAID3領域312からの読み出し、またはRAID3&4領域311からのリカバリ無しモードでの読み出し

次に、RAID3領域312からの読み出し、またはRAID3&4領域311からのリカバリ無しモードでの読み出しについて、図6のフローチャートを参照して説明する。

【0071】制御部15は、ホスト装置からの指示が、監視システムで適用されるビデオデータ等、短期間保存される大量データをRAID3領域312から読み出す場合、或いは通常のビデオデータ等、長期間保存されて繰り返し利用される大量データを直接ユーザに送信するためにRAID3&4領域311から読み出す場合、HDコントローラ部12内の各HDコントローラ120-1~120-nに対して、RAID4によるリカバリ機能無

しのリードのための命令(read命令)を発行する(ステップS31)。

【0072】HDコントローラ120-i(i=1~n)は、制御部15から与えられる命令(read命令)で要求されたHDD110-ij(jは1~mのいずれか)からのデータ読み出しを行い(ステップS32)、正常に読めたか否かをチェックする(ステップS33)。

【0073】もし、エラーがなかったならば、HDコントローラ120-iはステップS32で読み出したデータをバッファ部13内のバッファ130-iに格納し、制御部15に正常終了を通知する(ステップS34)。これに対してエラーがあったならば、HDコントローラ120-iはリカバリを行わず、制御部15にリードエラーを通知する(ステップS35)。

【0074】制御部15は、各HDコントローラ120-iからのエラー通知を監視しており(ステップS36)、エラーを通知したHDコントローラ120-iの数が0であるならば、バッファ部13内のバッファ130-1~130-nには、HDコントローラ120-1~120-nに要求したデータが格納されているものと判断してステップS37に進む。このステップS37では、制御部15はRAID機構14に対してRAID3の機能を起動させ、バッファ130-1~130-(n-1)のデータを合成させた後、その合成後のデータをホスト装置に出力する。

【0075】また、エラーを通知したHDコントローラ120-iの数が1であるならば、制御部15はステップS38に進む。このステップS38では、制御部15はRAID機構14に対してRAID3の機能を起動させ、エラーを通知した唯一のHDコントローラ120-iを除くHDコントローラ120-1~120-nに対応するバッファ130-1~130-nのデータから、エラーを通知したHDコントローラ120-iが読み出すべきデータをバッファ130-i内にリカバリさせ、バッファ130-1~130-(n-1)のデータを合成させた後、その合成後のデータをホスト装置に出力する。

【0076】また、エラーを通知したHDコントローラ120-iの数が2以上であるならば、制御部15はRAID機構14のRAID3機能を起動してもリカバリ不可能であるとして、その旨をホスト装置に通知してエラー終了する(ステップS39)。

(4) RAID3領域312への書き込み

次に、RAID3領域312への書き込みについて、図7のフローチャートを参照して説明する。

【0077】制御部15は、ホスト装置からの指示が、監視システムで適用されるビデオデータ等、短期間保存される大量データのRAID3領域312への書き込みを行うものである場合、まずRAID機構14に対してRAID3の機能を起動させる。これによりRAID機構14は、ホスト装置からのデータ(書き込みデータ)

をRAID3でストライピングして、そのパリティを生成しながら、そのストライピングしたデータ及びパリティを、バッファ部13内のそれぞれ対応するバッファ130-1~130-nに格納する(ステップS41)。

【0078】次に制御部15は、HDコントローラ部12内の各HDコントローラ120-1~120-nに対し、バッファ130-1~130-nに格納されたデータをRAID4でのパリティを生成せずに、RAID4グループ24-1~24-n内の目的とするHDD110-1j~110-nj (jは1~mのいずれか)に書き込むことを指示する(ステップS42)。

【0079】これを受けて各HDコントローラ120-i (i=1~n)は、バッファ130-iのデータをRAID4グループ24-i内の目的とするHDD110-ij (jは1~mのいずれか)に書き込む(ステップS43)。

(5) RAID4領域313からの読み出し  
次に、RAID4領域313からの読み出しについて、図8のフローチャートを参照して説明する。

【0080】制御部15は、ホスト装置からの指示が、静止画等の小量データのRAID4領域313からの読み出しである場合、目的とするHDD110-ij (iは1~nのいずれか、jは1~mのいずれか)を含むRAID4グループ24-iに対応するHDコントローラ部12内のHDコントローラ120-iに対して、RAID4によるリカバリ機能有りのリードのための命令(read命令)を発行する(ステップS51)。

【0081】HDコントローラ120-iは、制御部15から与えられる命令(read命令)で要求されたHDD110-ijからのデータ読み出しを行い(ステップS52)、正常に読めたか否かをチェックする(ステップS53)。

【0082】もし、エラーがなかったならば、HDコントローラ120-iはステップS52で読み出したデータをバッファ部13内のバッファ130-iに格納し、制御部15に正常終了を通知する(ステップS54)。

【0083】これに対し、エラーがあったならば、HDコントローラ120-iはRAID4によるリカバリを実行する(ステップS55)。即ちHDコントローラ120-iは、RAID4グループ24-i内のHDD110-ijを除く全てのHDD110-il~110-imを対象に、ステップS52で行ったのと同じ読み出しを行い、全て正常に読めたならば、リカバリ可能であるものと判断し、その読み出したデータの排他的論理和をとってHDD110-ijのデータをリカバリする。

【0084】HDコントローラ120-iは、RAID4によるリカバリができた場合(ステップS56)、リカバリしたデータをバッファ部13内のバッファ130-iに格納し、制御部15に正常終了を通知する(ステップS57)。これに対し、正常に読めなかったHDDが1

つでも存在するならば、HDコントローラ120-iはRAID4によるリカバリはできないものとして、制御部15に対して異常終了(リードエラー)を通知する(ステップS58)。

【0085】制御部15は、HDコントローラ120-iからの終了通知を監視しており(ステップS59)、正常終了が通知された場合には、RAID機構14に対してRAID3機能を起動させず、そのままバッファ130-iのデータをホスト装置に出力する(ステップS60)。これに対し、異常終了が通知された場合には、制御部15はリカバリ不可能あるとして、その旨をホスト装置に通知してエラー終了する(ステップS61)。

(6) RAID4領域313への書き込み  
次に、RAID4領域313への書き込みについて、図9のフローチャートを参照して説明する。

【0086】制御部15は、ホスト装置からの指示が、静止画等の小量データのRAID4領域313への書き込みである場合、まずホスト装置からのデータ(書き込みデータ)を、目的とするHDD110-ij (iは1~nのいずれか、jは1~mのいずれか)と接続されているHDコントローラ120-iに対応するバッファ130-iに格納する(ステップS71)。

【0087】次に制御部15は、HDコントローラ120-iに対し、バッファ130-iに格納されたデータを、RAID4でのパリティを生成しながら、RAID4グループ24-i内の目的とするHDD110-ijに書き込むことを指示する(ステップS72)。

【0088】これを受けてHDコントローラ120-iは、RAID4でのパリティを生成しながら、バッファ130-iのデータをRAID4グループ24-i内の目的とするHDD110-ijに書き込むと共に、生成したパリティをRAID4グループ24-i内のHDD110-imに書き込む(ステップS73)。なお、このRAID4でのパリティの生成は、1つのHDコントローラ120-iだけで行われる点を除けば、前記した(2)のRAID3&4領域311への書き込みにおけるステップS23でのパリティの生成と同様である。

【0089】さて、前記した図8のフローチャートに従うRAID4領域313からの読み出し(小量データの読み出し)では、目的とするHDD110-ijからのデータ読み出しが正常に行われなかったならば、そのリカバリのために、当該HDD110-ijを含むRAID4グループ24-i内の当該HDD110-ij以外の全てのHDD110-il~110-imを対象とする読み出し命令が発生する。このとき、ホスト装置から制御部15に対して新たなI/O要求が送られた場合、その要求は、上記のリカバリが終了するまで待たされる。しかし、新たなI/O要求の対象となるHDDが、例えばHDD110-ijを含むRAID3グループ23-jであるというように、並行してアクセス可能な場合でも、RAID4

による小量データの読み出しのためにその要求が待たされるのは無駄である。

【0090】そこで、このような不具合を解消するようにした、改良されたRAID4領域313からの読み出しにつき説明する。ここでは、制御部15からHDコントローラ120-iに対し、RAID4領域313からのデータ読み込み（小量データの読み込み）のためのリード命令を発行する際に、3つのモードの1つが指定可能となっている。この3つのモードは、(A) RAID4のリカバリ有りモード、(B) RAID4のリカバリ無しモード、(C) RAID4のリカバリモードである。以下、この3つのモードについて説明する。

(A) RAID4のリカバリ有りモード

このモードは、前記(5)のRAID4領域313からの読み出しと同様の動作を行うものであり、目的とするHDD110-ij (iは1～nのいずれか、jは1～mのいずれか)を含むRAID4グループ24-i内の全てのHDD110-il～110-imに対してリード命令が発生される可能性がある。

(B) RAID4のリカバリ無しモード

このモードでは、目的とするHDD110-ijのみにリード命令が発生する。

(C) RAID4のリカバリモード

このモードでは、目的とするHDD110-ijを含むRAID4グループ24-i内のHDD110-il～110-imのうち、目的とするHDD110-ijを除く全てのHDDへのリード命令のみが発生する。ここでは、目的とするHDD110-ijでのエラーの有無に無関係に、当該HDD110-ijを除くRAID4グループ24-i内の全HDDから読み込んだデータにより、目的とするHDD110-ijのデータをリカバリする。したがって、このモードでは、目的とするHDD110-ijへのアクセスが発生しないことが保証される。但し、この

(C)のモードでは読み込めなくて、上記(A)のモードでは読み込めるケースがある。これについて、図10を参照して説明する。

【0091】まず、RAID4グループ24-i内のHDD110-il～110-imのうち、図10に示すようなHDD110-il内の領域100-lからの読み込みがホスト装置から制御部15に対して指示されたものとする。ここで領域（読み込み対象領域）100-lは3セクタ分のサイズがあり、先頭セクタのみにエラー箇所があるものとする。また、HDD110-il内の領域100-lに対応するHDD110-i2内の領域100-2には、最終セクタのみにエラー箇所があり、HDD110-il～110-imのうち、HDD110-il、110-i2を除く残りのHDDの対応する領域には、エラー箇所は無いものとする。またHDDからの読み込みは、最小の読み込み単位であるセクタ単位で行われるものとする。

【0092】このような場合、上記(C)のリカバリモ

ードでHDD110-il内の領域100-lのデータを読み込もうとすると、領域100-lの先頭セクタと2番目のセクタについては、HDD110-i2～110-imの対応するセクタにエラー箇所が無いため、その対応するセクタの排他的論理和をとることで取得（リカバリ）可能である、これに対し、領域100-lの最後のセクタについては、HDD110-i2の対応するセクタにエラー箇所があるためエラーが発生し、失敗に終わる。勿論、HDD110-ilを除く残りのHDD110-i2～110-imの対応する領域にエラー箇所が無いならば、このHDD110-i2～110-imの対応する領域からの読み込みを行って、その排他的論理和をとることで、目的とするHDD110-ilの領域100-l内のデータに相当するデータを取得（リカバリ）することが可能である。

【0093】一方、上記(A)のリカバリ有りモードでHDD110-ilの領域100-lのデータを読み込む場合には、まず領域100-lの先頭セクタを読み込もうとしてエラーが発生するが、HDD110-i2～110-imの対応するセクタにエラー箇所が無いため、その対応するセクタの排他的論理和をとることでリカバリ可能である。領域100-lの後続のセクタ（2番目と3番目のセクタ）については、エラー箇所が無く、HDD110-ilからの読み込みが正常に行われるため、正常に終了する。

【0094】さて、本実施形態では、図1中の制御部15内に、RAID4領域313への入出力のための待ち行列であるキュー（キューエントリ列）を、各RAID4グループ（ここでは、RAID4グループ24-1～24-n）毎に有している。図11に、1つのRAID4グループに対応するキュー（以下、キュー#3と称する）の一例を示す。

【0095】図11において、ヘッダ部111に保持されたポインタ112で指定されるキューエントリ（先頭のキューエントリ）113-1は、リードであるか或いはライトであるかを示すフラグ（R/Wフラグ）1131、リード或いはライトの対象となるHDD、即ち目的とするHDD（ディスク装置）の識別子（装置識別子）1132、リード／ライトデータを一時的に格納するバッファ130-i内の格納先頭番地1133、シークアドレス（装置識別子の示すHDD内のRAID4領域313のシークアドレス）1134、転送長1135、モード識別子1136、他のI/Oとの干渉が無い場合にかかる時間（想定される実行時間）の上限値1137、及び次のキューエントリを指す次エントリポインタ1138を有している。モード識別子1136は、R/Wフラグ1131がリードを示す場合には、RAID4のリカバリ無しモード、RAID4のリカバリモード、及びRAID4のリカバリ有りモードのいずれか1つを示す。また、モード識別子1136は、R/Wフラグ11

31がライトを示す場合には、常にパリティを生成するモードを示す。キューエントリ113-1につながる他のキューエントリ113-2…も、当該キューエントリ113-1と同様のデータ構造をなす。

【0096】ここで、先頭のキューエントリ113-1に設定されているI/O要求の実行について、当該キューエントリ113-1が、ホスト装置から制御部15に対してRAID4グループ24-i中のHDD110-ilの領域100-lからのRAID4での読み込みが指示された結果生成された場合を例に、図12のフローチャートを参照して説明する。このときキューエントリ113-1中のR/Wフラグ1131はリードを、装置識別子1132はHDD110-ilを、シークアドレス1134はHDD110-il内の領域100-lの先頭アドレスを、転送長1133は3セクタを、モード識別子1136はRAID4のリカバリ無しモードを示しているものとする。

【0097】まず制御部15は、キューエントリ113-1の内容に従って、RAID4のリカバリ無しモードでの読み込みを行うようにHDコントローラ部12内のHDコントローラ120-iを制御する（ステップS81）。

【0098】この場合、HDコントローラ120-iはHDD110-ilの指定領域100-lからのデータ読み込みのみを行う。もし、読み込みエラーとなった場合には、制御部15は実行中のキューエントリ113-1のモード識別子をRAID4のリカバリモードに更新すると共に、他のI/Oとの干渉が無い場合にかかる時間の上限値1137を当該RAID4のリカバリモードに合わせて更新し、当該キューエントリ113-1をつなぎ直す再スケジューリングを行って、機会を待つ。この上限値1137の更新については後述する。

【0099】制御部15は、上記更新後のキューエントリ113-1が実行可能になると、当該エントリ113-1の内容に従って、RAID4のリカバリモードでの読み込みを行うようにHDコントローラ部12内のHDコントローラ120-iを制御する（ステップS82）。

【0100】この場合、HDコントローラ120-iはHDD110-ilの指定領域100-lに対応する他のHDD110-i2～HDD110-imの領域からのデータ読み込みを行い、正常に読み込んだら、その排他的論理和をとることで、指定領域100-lのデータに相当するデータを取得（リカバリ）する。一方、1つでも読み込みエラーが発生したならば、制御部15は実行中のキューエントリ113-1のモード識別子をRAID4のリカバリ有りモードに更新すると共に、他のI/Oとの干渉が無い場合にかかる時間の上限値1137を当該RAID4のリカバリ有りモードに合わせて更新し、当該キューエントリ113-1をつなぎ直して、機会を待つ。

【0101】ここで、上記RAID4のリカバリ無しモードでの読み込みと、上記RAID4のリカバリモードでの読み込みの期間、その読み込みの対象となっているHDDを除くHDDへのI/O要求は実行可能である。

【0102】制御部15は、上記更新後のキューエントリ113-1が実行可能になると、当該エントリ113-1の内容に従って、RAID4のリカバリ有りモードでの読み込みを行うようにHDコントローラ部12内のHDコントローラ120-iを制御する（ステップS83）。

【0103】この場合、HDコントローラ120-iは、まずHDD110-ilの指定領域100-lからのデータ読み込みを行い、読み込みエラーが発生したならば、HDD110-ilを除く残りのHDD110-i2～110-imの対応する領域からの読み込みを行い、正常に読み込んだら、その排他的論理和をとることで、目的とするHDD110-ilの領域100-l内のデータに相当するデータをリカバリする動作を、例えば読み込みの最小単位（ここではセクタ単位）で繰り返す。

【0104】このように本実施形態においては、RAID4での読み込みを最初から（他のI/O要求が待たされる）リカバリ有りモードで行うのではなくて、まず

（目的のHDD以外を対象とするI/O要求の並列実行が可能）RAID4のリカバリ無しモードで読み込みを行い、読み込みエラーが発生したならば、（目的のHDDを対象とする別のI/O要求の並列実行が可能）RAID4のリカバリモードで読み込みを行い、それでも読み込みエラーが発生した場合に、RAID4のリカバリ有りモードでの読み込みを行うようにしている。

【0105】即ち本実施形態においては、RAID4での小量データの読み込みを、ビデオデータ等の大量データの読み込みのスケジューリングに極力影響を与えないような方法を優先的に適用して実行し、動作するHDDを細かく制御しているため、大量データの読み込みのスケジューリングに影響を与えないようにしながら小量データの読み込みを行う機会を増加させることができる。

【0106】なお、制御部15はスケジューリングの際、他のI/O要求、特にRAID3&4領域311、RAID3領域312への（大量データの）I/O要求の混み具合が少ない場合には、ステップS81、S82を省略する。

【0107】さて、先に述べた図4のフローチャートに従うRAID3&4領域311からのリカバリ有りモードでの読み出しでは、各HDコントローラ120-i（i=1～n）は、制御部15から与えられる命令（read命令）で要求されたHDD110-ij（jは1～mのいずれか）からのデータ読み出しを行い、正常に読めなかったら、そのHDコントローラ120-iはRAID4グループ24-i内のHDD110-i1～110-imをアクセスしてRAID4でのリカバリ処理を行う。しかし、1台のHDコントローラ120-iのみがリカバリ処理を

行う場合には、バッファ130-l~130-mの内容をもとにRAID機構14にてRAID3でのリカバリが可能なることから、その時間(RAID4でのリカバリ処理の時間)が無駄となる。

【0108】そこで、このような不具合を解消するようにした、改良されたRAID3&4領域311からのリカバリ有りモードでの読み出しにつき説明する。ここでは、前記したRAID4領域313への入出力を管理するためのキュー#3とは別に、RAID3&4領域311及びRAID3領域312への入出力を管理するための、次に述べる2種類のキュー(キューエントリ列)を制御部15内に有している。

【0109】図13は、RAID3&4領域311及びRAID3領域312への入出力のための通常のI/O要求のキュー(以下、キュー#1と称する)の一例を示す。このキュー#1は、RAID3のグループ(ここでは、RAID3グループ23-1~23-m)の数だけ設けられる。

【0110】図13において、ヘッダ部131に保持されたポインタ132で指定されるキューエントリ(先頭のキューエントリ)133-1は、リードであるか或いはライトであるかを示すフラグ(R/Wフラグ)133-1、リード/ライトデータを一時的に格納するバッファ130-i内の格納先先頭番地1133、シークアドレス(RAID3グループ23-i内のシークアドレス)1134、転送長1135、モードフラグ1136、及び次のキューエントリを指す次エントリポインタ1138を有している。キューエントリ133-1につながる他のキューエントリ133-2...も、当該キューエントリ133-1と同様のデータ構造をなす。

【0111】モードフラグ1336は、R/Wフラグ1331がリードを示す場合には、元のI/O要求にRAID4のリカバリが必要か否かを示す。ここでは、RAID3&4領域311からのリカバリ有りモードでの読み出しの場合のみ、元のI/O要求にRAID4のリカバリが必要となる。また、モードフラグ1336は、R/Wフラグ1331がライトを示す場合にはRAID4のパリティ生成が必要か否かを示す。ここでは、RAID3&4領域311への書き込みの場合のみ、パリティ生成が必要となる。

【0112】図14は、RAID3&4領域311に対するRAID4のリカバリ有りモードの読み込み、またはRAID4のパリティ生成(パリティ生成・書き込み)の要求のキュー(以下、キュー#2と称する)の一例を示す。このキュー#2は、RAID4のグループ(ここでは、RAID4グループ24-1~24-n)の数だけ設けられる。

【0113】図14において、ヘッダ部141に保持されたポインタ142で指定されるキューエントリ(先頭のキューエントリ)143-1は、リード(RAID3&

4領域311に対するRAID4のリカバリ有りモードの読み込み)であるか或いはライト(RAID4のパリティ生成が必要な書き込み)であるかを示すフラグ(R/Wフラグ)1431、リード或いはライトの対象となるHDD、即ち目的とするHDD(ディスク装置)の識別子(装置識別子)1432、リード/ライトデータを一時的に格納するバッファ130-i内の格納先先頭番地1433、シークアドレス(装置識別子の示すHDD内のシークアドレス)1434、転送長1435、他のI/Oとの干渉が無い場合にかかる時間の上限値1437、及び次のキューエントリを指す次エントリポインタ1438を有している。キューエントリ143-1につながる他のキューエントリ143-2...も、当該キューエントリ143-1と同様のデータ構造をなす。

【0114】ここで、キュー#1の先頭のキューエントリ133-1に設定されているI/O要求の実行について、当該キューエントリ133-1が、ホスト装置から制御部15に対してRAID3&4領域311からのリカバリ有りモードでの読み込みが指示された結果生成された場合を例に、図15のフローチャートを参照して説明する。このときキューエントリ133-1中のR/Wフラグ1331はリードを、モードフラグ1336は元のI/O要求にRAID4のリカバリが必要であることを示しているものとする。

【0115】まず制御部15は、キュー#1の先頭エントリであるキューエントリ133-1の内容に従って、HDコントローラ部12内の各HDコントローラ120-i(i=1~n)に対してRAID4のリカバリ機能無しのリードのための命令(read命令)を発行し、結果を受け取る(ステップS91)。

【0116】HDコントローラ120-iは制御部15からの読み込み命令で要求されたHDD110-ij(jは1~mのいずれか)からの読み出しのみを行い、正常に読み込んだなら、その読み込んだデータをバッファ130-iに格納し、制御部15に正常終了を通知する。これに対し、HDD110-ijから正常に読み込めなかったならば、HDコントローラ120-iは制御部15にエラーを通知する。

【0117】制御部15は、上記ステップS91で各HDコントローラ120-iからの終了通知を受け取ると、エラーを通知したHDコントローラ120-iの数を調べる(ステップS92)。

【0118】もし、エラーを通知したHDコントローラ120-iの数が0であるならば、制御部15は、バッファ部13内のバッファ130-l~130-nには、HDコントローラ120-1~120-nに要求したデータが格納されているものと判断してステップS93に進む。このステップS93では、制御部15はRAID機構14に対してRAID3の機能を起動させ、バッファ130-l~130-(n-1)のデータを合成させた後、その合成後の

データをホスト装置に出力する。

【0119】また、エラーを通知したHDコントローラ120-iの数が1であるならば、制御部15はステップS94に進む。このステップS94では、制御部15はRAID機構14に対してRAID3の機能を起動させ、エラーを通知した唯一のHDコントローラ120-iを除くHDコントローラ120-1~120-nに対応するバッファ130-1~130-nのデータから、エラーを通知したHDコントローラ120-iが読み出すべきデータをバッファ130-i内にリカバリさせ、バッファ130-1~130-(n-1)のデータを合成させた後、その合成後のデータをホスト装置に出力する。

【0120】また、エラーを通知したHDコントローラ120-iの数が2以上であるならば（このように、2つ以上のHDコントローラからエラーが通知される確率は、1つのHDコントローラがエラーを通知する確率よりもはるかに小さい）、制御部15はキューエントリ133-1に対応する図14に示したようなキューエントリ143-1を生成してキュー#2につなぎ、そのキューエントリ143-1の内容に従って、エラーを通知した2つ以上のHDコントローラ120-iに対し、RAID4のリカバリ機能有りのリードのための命令（read命令）を発行し、結果を受け取る（ステップS95）。

【0121】制御部15からRAID4のリカバリ機能有りのリードのための命令が与えられた2つ以上のHDコントローラ120-iは、制御部15からの読み込み命令で要求されたHDD110-ij（jは1~mのいずれか）からの読み出しを行い、エラーがあったなら、前記した図4中のステップS5と同様にRAID4によるリカバリを実行する。そしてHDコントローラ120-iは、リカバリできたなら、制御部15に対して正常終了を通知し、リカバリできなかったなら、制御部15に対してエラーを通知する。

【0122】制御部15は、上記ステップS95で各HDコントローラ120-iからの終了通知を受け取ると、エラーを通知したHDコントローラ120-iの数を調べる（ステップS96）。

【0123】もし、エラーを通知したHDコントローラ120-iの数が0であるならば、制御部15は、バッファ部13内のバッファ130-1~130-nには、HDコントローラ120-1~120-nに要求したデータが格納されているものと判断して上記ステップS93に進む。このステップS93では、制御部15はRAID機構14に対してRAID3の機能を起動させ、バッファ130-1~130-(n-1)のデータを合成させた後、その合成後のデータをホスト装置に出力する。

【0124】また、エラーを通知したHDコントローラ120-iの数が1であるならば、制御部15は上記ステップS94に進む。このステップS94では、制御部15はRAID機構14に対してRAID3の機能を起動

させ、エラーを通知した唯一のHDコントローラ120-iを除くHDコントローラ120-1~120-nに対応するバッファ130-1~130-nのデータから、エラーを通知したHDコントローラ120-iが読み出すべきデータをバッファ130-i内にリカバリさせ、バッファ130-1~130-(n-1)のデータを合成させた後、その合成後のデータをホスト装置に出力する。

【0125】このように本実施形態においては、1つのHDコントローラ120-iだけがエラーを通知した場合には、エラーを通知した唯一のHDコントローラ120-iを除くHDコントローラ120-1~120-nに対応するバッファ130-1~130-nのデータから、エラーを通知したHDコントローラ120-iが読み出すべきデータをバッファ130-i内にリカバリできる。即ち、リトライ時間がかからずに目的データをバッファ130-iに読み出せる。しかも、2つ以上のHDコントローラからエラーが通知される確率は極めて小さいが、そのような場合でも、遅くはなるが、目的のデータを読み出すことが可能となる。したがって、システム全体としては、データ読み出し時間が短縮される。

【0126】さて、先に述べた図5のフローチャートに従うRAID3&4領域311への書き込みでは、RAID4でのパリティの生成のために、例えばRAID4グループ24-iを例にとると、当該グループ24-i内のHDD110-i1~110-i(m-1)の対応する領域からデータを読み出し、その排他的論理和をとってパリティを生成しなければならない。このため、RAID3&4領域311への書き込みに多大な時間を要し、例えばビデオカメラからのビデオ信号をデジタル化し、更にエンコードしたもの（ビデオエンコーダの出力）をリアルタイムでディスクアレイ装置のRAID3&4領域311に格納することは困難である。

【0127】そこで、このような不具合を解消するようにした、改良されたRAID3&4領域311への書き込みについて、図16のフローチャートを参照して説明する。

【0128】まず制御部15は、ホスト装置からRAID3&4領域311への書き込みが指示された場合、前記キュー#1につなぐべき（登録すべき）キューエントリを生成し、当該キュー#1につなぐ（ステップS101）。ここでは、図13に示したキューエントリ133-1が生成されてキュー#1につながれたものとする。この場合、キューエントリ133-1のR/Wフラグ1331はライトを示し、モードフラグ1336は、RAID4のパリティ生成が必要であることを示す。

【0129】制御部15は、キューエントリ133-1に従ってRAID機構14に対してRAID3の機能を起動させることでホスト装置からのデータとそのパリティをバッファ部13内のそれぞれ対応するバッファ130-1~130-nに格納させると共に、各HDコントローラ



120-1~120-mを制御することで、バッファ130-1~130-n内のデータ（ホスト装置からのデータまたはそのパリティ）を、指定のRAID3グループ23-j（jは1~m-1のいずれか）内のHDD110-1j~HDD110-njに書き込むRAID3による書き込みを行わせる（ステップS102）。

【0130】制御部15は、キューエントリ133-1に従うRAID3による書き込みが正常終了したならば、当該キューエントリ133-1をもとに前記したキュー#2につなぐべきキューエントリを生成し、当該キュー#2につなぐ（ステップS103）。この処理は、RAID4グループ毎に行われる。ここでは、図14に示したキューエントリ143-1が生成されてキュー#2につなぐられたものとする。この場合、キューエントリ143-1のR/Wフラグ1331はライトを示す。

【0131】制御部15は、生成したキューエントリ143-1をキュー#2につなぐと、RAID4でのパリティ生成と書き込みが行われなくとも、ホスト装置に正常終了を通知する（ステップS104）。

【0132】一方、RAID4グループ毎のキュー#2につなぐられた各キューエントリ143-1は後で取り出されて実行され、RAID4でのパリティが生成される（ステップS105）。このRAID4グループ毎に生成されたパリティは、そのグループ内のHDD110-i<sub>m</sub>（i=1~n）に書き込まれる。

【0133】このように本実施形態では、RAID3&4領域311への書き込みを、RAID3によるパリティ生成を行いながらのデータ及びパリティ書き込みと、RAID4によるパリティ生成・書き込みとの2段階で行い、短時間で処理できる前者の書き込みが終了（正常終了）した段階で、ホスト装置に対して終了（正常終了）を通知する構成としている。このため、RAID3&4領域311への書き込み時の見かけ上の性能向上が可能となり、リアルタイムエンコードの出力をそのままディスクアレイ装置に格納するような場合にも、必要な性能が確保できる。

【0134】なお、本実施形態では、上記した後者の書き込み（RAID4によるパリティ生成・書き込み）が終了するまでの間は、該当するデータはRAID3で保護されているだけであるが、これだけでもある程度の保証はでき、しかも一時的な状態であることから、問題はない。

【0135】但し、上記ステップS103では、生成したキューエントリ143-1を、同一キュー#2内のどの「RAID4のリカバリ有りモードの読み込み」を要求するキューエントリよりも前方につなぐ必要がある。その理由は、もし、RAID4のリカバリ有りモードの読み込み→他のRAID3グループへの書き込み→RAID4のリカバリ有りモードの読み込みのリトライ→パリティ生成・書き込みという順序になると、誤ったデータ

を読み込むことになるので、そのような順序になるのを防ぐためである。

【0136】また、RAID4のリカバリ有りモードの読み込み→同じRAID3グループへの書き込み→RAID4のリカバリ有りモードの読み込みのリトライ→パリティ生成・書き込みという順序になると、読み込み命令（read命令）発行時より後の時点のデータを読み込むことになるが、これを避けるのはホスト装置側の責任であるとする。

【0137】次に、図1のディスクアレイ装置全体のI/O要求に対するスケジューリングについて、図17及び図18のフローチャートを参照して説明する。まず本実施形態では、制御部15内の不揮発性メモリ150の所定番地の領域に、前記第1の実施形態で述べたレジスタ151と同様に、次に述べる3つのレジスタ、即ちtレジスタ152、Tレジスタ153及びAレジスタ154が割り当てらる。

【0138】tレジスタ152は、制御部15が使用する変数tを記憶するためのもので、Tレジスタ153の値（T）とI/O処理に実際に要した時間との差分を蓄積していくために使用される。

【0139】Tレジスタ153は、システム立ち上げ時にホスト装置からの制御命令で指定される値Tを記憶しておくためのものである。本実施形態では、キュー#1への登録対象となるI/O要求は、その転送長が予め定められた一定の長さ以上の場合には、一定長を最大の転送長とする分割した要求として扱われ、複数のキューエントリが生成されてキュー#1につながれるようになっている。その理由は、キュー#1につながれる各エントリの内容（の示すI/O要求）を実行するのに想定される所要時間をほぼ一定とするためである。この値に予め定められた余裕時間を加えた時間が上記Tとしてホスト装置から与えられてTレジスタ153に設定されることになる。

【0140】Aレジスタ154は、システム立ち上げ時にホスト装置からの制御命令で指定される値Aを記憶しておくためのものである。前記したように、キュー#2、#3では、キューエントリをつなぐ際に、必要なI/Oの回数、転送長等から所要時間の上限値を見積もり、そのエントリ中に、他のI/Oとの干渉が無い場合にかかる時間の上限値1137、1437として設定される。Aレジスタ154に設定される値Aには、キュー#2、#3に登録されるI/O要求の実行に要する時間の上限値（他のI/Oとの干渉が無い場合にかかる時間の上限値1137、1437）及びTレジスタ153の内容（T）より大きい値が用いられる。

【0141】さて、図1のディスクアレイ装置では、システム立ち上げ時に、tレジスタ152に初期値0がセットされる（ステップS111）。やがてホスト装置から制御部15に対して何らかのI/O要求が与えられた

ならば、制御部15はそのI/O要求を取り込んで、そのI/O要求に適合したキュー用のキューエントリを生成し、対応するキュー（キュー#1～#3のいずれか）につなぐ（ステップS112）。

【0142】次に制御部15は、レジスタ152の値 $t$ とレジスタ154の値 $A$ の大きさを比較し、 $t < A$ の条件が成立するか否かをチェックする（ステップS113）。この時点では、上記ステップS111の処理で $t = 0$ となっているため、 $t < A$ の条件が成立する。

【0143】すると制御部15は、各RAID3グループ23-1～23- $m$ 毎に用意されるキュー#1から、RAID3のグループが重複しないように、キューエントリを取り出し、スケジュールする（ステップS114）。

【0144】次に制御部15は、各RAID4グループ24-1～24- $n$ 毎に用意されるキュー#2から、所要時間の上限値（他のI/Oとの干渉が無い場合にかかる時間の上限値1437）がレジスタ153の値 $T$ より小さく且つRAID4のグループが重複しないようなキューエントリを取り出し、スケジュールする（ステップS115）。

【0145】次に制御部15は、各RAID4グループ24-1～24- $n$ 毎に用意されるキュー#3から、所要時間の上限値（他のI/Oとの干渉が無い場合にかかる時間の上限値1137）がレジスタ153の値 $T$ より小さく且つ上記ステップS114、S115でスケジュールしたI/OとHDDが重複しないようなキューエントリを取り出し、スケジュールする（ステップS116）。

【0146】次に制御部15は、ステップS113～S116で少なくとも1つスケジュールしたか否かをチェックし（ステップS117）、少なくとも1つスケジュールしたときには、レジスタ152の値 $t$ を、その値 $t$ にレジスタ153の値 $T$ を加えた $t + T$ （ここでは $T$ ）に更新する（ステップS118）。

【0147】次に制御部15は、ステップS113～S116でスケジュールしたI/O要求を各HDDコントローラ120- $i$ に発行し、全て終了するまで待つ（ステップS119）。そして制御部15は、ステップS119で要求したI/Oが全て終了すると、レジスタ152の値 $t$ を、その値 $t$ （ここでは $t = T$ ）から当該I/Oの経過時間（所用時間）を差し引いた値に更新し（ステップS120）、ステップS112に戻る。ここで、 $T > I/O$ の経過時間であることから、上記ステップS118～S120の実行後のレジスタ152の値 $t$ は増加する。また、実際に経過する時間は $T$ 未満である。

【0148】以上により、ステップS118～S120を繰り返し実行すると、やがて、ステップS113の条件（ $t < A$ ）が成立しなくなる。また、ステップS113～S116で1つもスケジュールできなかった場合には、レジスタ152の値 $t$ がレジスタ154の値 $A$

に更新されて（ステップS121）、ステップS112に戻るため、直ちにステップS113の条件（ $t < A$ ）が成立しなくなる。

【0149】このように本実施形態においては、ステップS112からステップS120までのループは、いずれ終了する。さて、ステップS113の条件（ $t < A$ ）が成立しなくなると、制御部15はステップS122以降の処理に進む。即ち制御部15は、まず全てのキュー#2から、RAID4のグループが重複しないように、キューエントリを取り出し、スケジュールする（ステップS122）。

【0150】次に制御部15は、全てのキュー#3から、上記ステップS122でスケジュールしたI/OとHDDが重複しないようなキューエントリを取り出し、スケジュールする（ステップS123）。

【0151】次に制御部15は、ステップS122、S123で少なくとも1つスケジュールしたか否かをチェックし（ステップS124）、少なくとも1つスケジュールしたときには、ステップS122、S123でスケジュールしたI/O要求を各HDDコントローラ120- $i$ に発行し、全て終了するまで待つ（ステップS125）。そして制御部15は、ステップS125で要求したI/Oが全て終了すると、レジスタ152の値 $t$ を初期値0に更新し（ステップS126）、ステップS112に戻る。また、ステップS122、S123で1つもスケジュールできなかった場合には、制御部15はステップS125をスキップしてステップS126に進み、レジスタ152の値 $t$ を初期値0に更新する。

【0152】このように本実施形態においては、ステップS122以降の処理が実行されると、最終的にステップS126でレジスタ152の値 $t$ が0となり、上記ステップS113の条件（ $t < A$ ）が成立する。したがって、ステップS122以降の処理が続けて実行されることはない。このステップS122以降の処理を実行した場合、実際の経過時間は、 $A$ 未満である。

【0153】以上の説明から明らかなよう、本実施形態におけるスケジューリングでは、キュー#1が空きの状態で新しいエントリが登録された場合には、そのエントリ（の示すI/O要求）は（ $A + T$ ）時間以内にスケジュールされる。また1つのキュー#1に例えば $p$ 個のエントリが繋がっている状態で、新しいエントリが登録された場合には、先頭のエントリがスケジュールされるまでの最長時間は（ $A + T$ ）であり、 $p$ 個のエントリのI/Oに要する時間の上限は $pT$ であることから、新たに登録されたエントリ（の示すI/O要求）は（ $A + (p + 1)T$ ）時間以内にスケジュールされる。即ち、キュー#1へのI/Oの頭出し時間が保証できる。

【0154】また、本実施形態におけるスケジューリングでは、キュー#1が空にならないようにホスト装置から制御することにより、 $T$ 時間に1回の割合でI/O要求

を実行することができる。即ち、キュー#1へのI/Oのレートが保証できる。但し、リカバリ有りモードのI/O要求を発行した場合で、キュー#2につなぎかえられた場合には、その終了の通知は遅れる。

【0155】また、本実施形態におけるスケジューリングでは、キュー#1にエントリがなくなる状態でも、キュー#2、#3のエントリにはスケジュールされる機会が訪れる。

【0156】このように本実施形態によれば、次のような効果を得ることができる。

(1) RAID3&4領域311及びRAID3領域312へのI/O(入出力)において、I/O要求の発行及び終了を監視し、キューの状態をホスト装置から想定することにより、今、発行しようとするI/O要求に対する最悪の実行開始時点、RAID4領域313への、(その後発行されるI/O要求も含めた)I/O状況に無関係に予測できる。

(2) RAID3&4領域311及びRAID3領域312へのI/Oにおいて、ホスト装置からI/O要求を十分に先出しすることにより、該当するRAID3グループのHDDのI/Oのレートを保証することができる。但し、RAID3&4領域311からの読み込みの場合に、RAID4のリカバリ有りモードのI/O要求を発行した場合、その終了通知は遅れることがあり得る。また、各I/Oの終了は一定間隔ではなく、前倒しで終了することがあり得る。

(3) RAID3&4領域311及びRAID3領域312へのI/Oのための要求が頻繁に発行される状況でも、RAID4領域313へのI/Oは必ず「暇をみて」行われる。この場合、(要求するデータ転送長の上限を小さくするようにするなどして)I/Oの所要時間の上限値を極力減少させるように、ホスト装置から制御する必要があるが、ホスト装置から制御コマンドを発行することで、(Aの値を小さくし)この頻度を増加させることも可能である。但し、上記(1)の予測時間を減少させる効果を生むので、ホスト装置側では、慎重にAの値を設定する必要がある。

【0157】以上に述べたように、本実施形態によれば、ビデオデータのような大量のデータの入出力については、一定の入出力レートを保証し、且つ余裕があれば、静止画などの小量データの入出力も並行して実行されるディスクアレイ装置を実現できる。

【0158】以上に述べた実施形態では、図1中のHDD群110におけるy方向(縦方向)の配列であるHDD110-11~110-n1, ..., 110-1m~110-nmは、図2に示したように、それぞれRAID3グループ23-1, ..., 23-mを構成しているものとして説明したが、これに限るものではない。例えば、上記したy方向(縦方向)の配列であるHDD110-11~110-n1, ..., 110-1m~110-nmも、図19に示すよう

に、RAID4グループ25-1, ..., 25-mとするようにしても構わない。この構成では、図1中の制御部15に、RAID3の実行機能に代えてRAID4の実行機能を持たせる必要がある点を除けば、前記第1の実施形態と同様である。

【0159】なお、図19の構成では、1つのHDコントローラ120-i(i=1~n)下に1つRAID4グループ24-iが接続されているが、全てのHDコントローラ120-i(i=1~n)に同数の複数のグループが接続される構成であっても構わない。

【0160】図19のような構成とした場合、I/Oエラーのリカバリが必要になったとき、RAID機構14によるバッファ部13のアクセス頻度が増大するものの、(この構成ではRAID4&4領域と呼ぶべき)RAID3&4領域311に静止画等の小量データを配置することが可能となり、小量データに2重のRAID保護を行うことが可能となる。

【0161】

【発明の効果】以上詳述したように本発明によれば、同一ディスク装置上で複数のディスクアレイ方式を可能とすると共に、2重のディスクアレイ保護を可能とし、信頼性の向上を図ると共に、異種データの混在を可能とし、しかも各データ種類・用途に応じた最適なアクセスが行える。

#### 【図面の簡単な説明】

【図1】本発明の一実施形態に係るマルチメディアサーバ用ディスクアレイ装置の構成を示すブロック図。

【図2】図1中のHDD群11におけるRAID3及びRAID4のグループの構成例を示す図。

【図3】図1中の各HDD110-ij(i=1~n, j=1~m)に装着されるディスクの領域分割例を示す図。

【図4】同実施形態におけるRAID3&4領域からのリカバリ有りモードでの読み出しを説明するためのフローチャート。

【図5】同実施形態におけるRAID3&4領域への書き込みを説明するためのフローチャート。

【図6】同実施形態におけるRAID3領域からの読み出しと、RAID3&4領域からのリカバリ無しモードでの読み出しを説明するためのフローチャート。

【図7】同実施形態におけるRAID3領域への書き込みを説明するためのフローチャート。

【図8】同実施形態におけるRAID4領域からの読み出しを説明するためのフローチャート。

【図9】同実施形態におけるRAID4領域への書き込みを説明するためのフローチャート。

【図10】図8に示したRAID4領域からの読み出しを改良するために用意される3つのモードのうちの、RAID4のリカバリ有りモードとRAID4のリカバリ無しモードとの違いを説明するための図。

【図11】上記3つのモードを適用する場合のRAID4領域への入出力のためのキュー（#3）の構造例を示す図。

【図12】図8に示したRAID4領域からの読み出しを改良したフローチャート。

【図13】図4に示したRAID3&4領域からのリカバリ有りモードでの読み出しを改良するために適用される、RAID3&4領域及びRAID3領域への入出力を管理するためのキュー（#1）の構造例を示す図。

【図14】図4に示したRAID3&4領域からのリカバリ有りモードでの読み出しを改良するために適用される、当該読み出し、またはRAID4のパリティ生成を要求するためのキュー（#2）の構造例を示す図。

【図15】図4に示したRAID3&4領域からのリカバリ有りモードでの読み出しを改良したフローチャート。

【図16】図5に示したRAID3&4領域への書き込みを改良したフローチャート。

【図17】同実施形態におけるI/O要求に対するスケジューリングを説明するためのフローチャートの一部を示す図。

【図18】同実施形態におけるI/O要求に対するスケジューリングを説明するためのフローチャートの残りを示す図。

【図19】図2の構成の変形例を示す図。

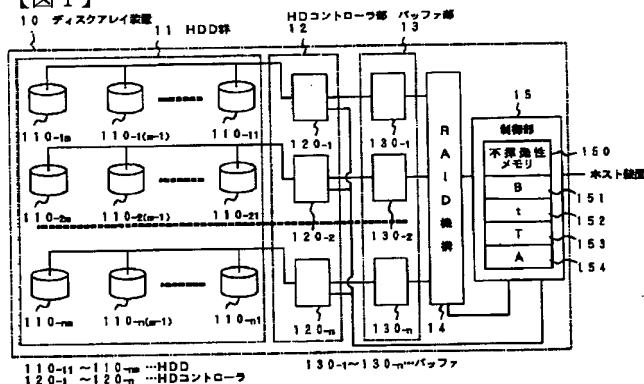
【図20】従来のディスクアレイ装置のブロック構成図。

【図21】従来のディスクアレイ装置のブロック構成図。

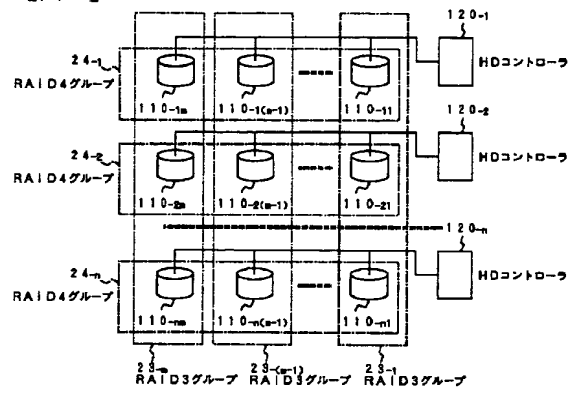
【符号の説明】

- 10…ディスクアレイ装置、
- 11…HDD群（ディスク装置の群）、
- 12…HDコントローラ部、
- 13…バッファ部、
- 14…RAID機構（入出力手段）、
- 15…制御部、
- 23-1～23-m…RAID3グループ（第1のディスクアレイグループ）、
- 24-1～24-n, 25-1～25-m…RAID4グループ（第2のディスクアレイグループ）、
- 110-11～110-nm, 110-ij…HDD（ディスク装置）、
- 120-1～120-n…HDコントローラ（ディスクコントローラ）、
- 130-1～130-n…バッファ、
- 150…不揮発性メモリ、
- 151…Bレジスタ、
- 152…tレジスタ（第3の時間記憶手段）、
- 153…Tレジスタ（第1の時間記憶手段）、
- 154…Aレジスタ（第2の時間記憶手段）、
- 311…RAID3&4領域（第1の領域）、
- 312…RAID3領域（第2の領域）、
- 313…RAID4領域（第3の領域）。

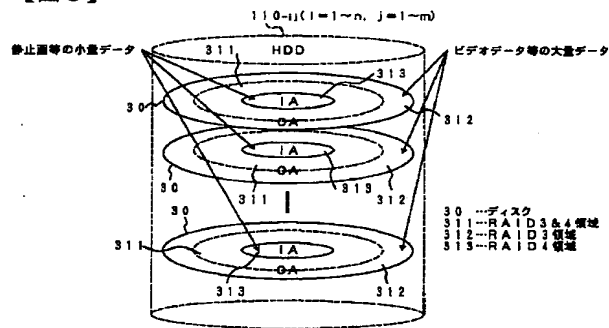
【図1】



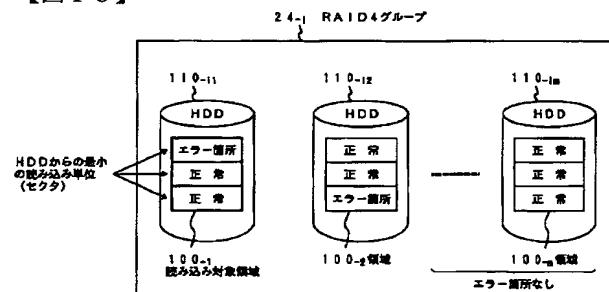
【図2】



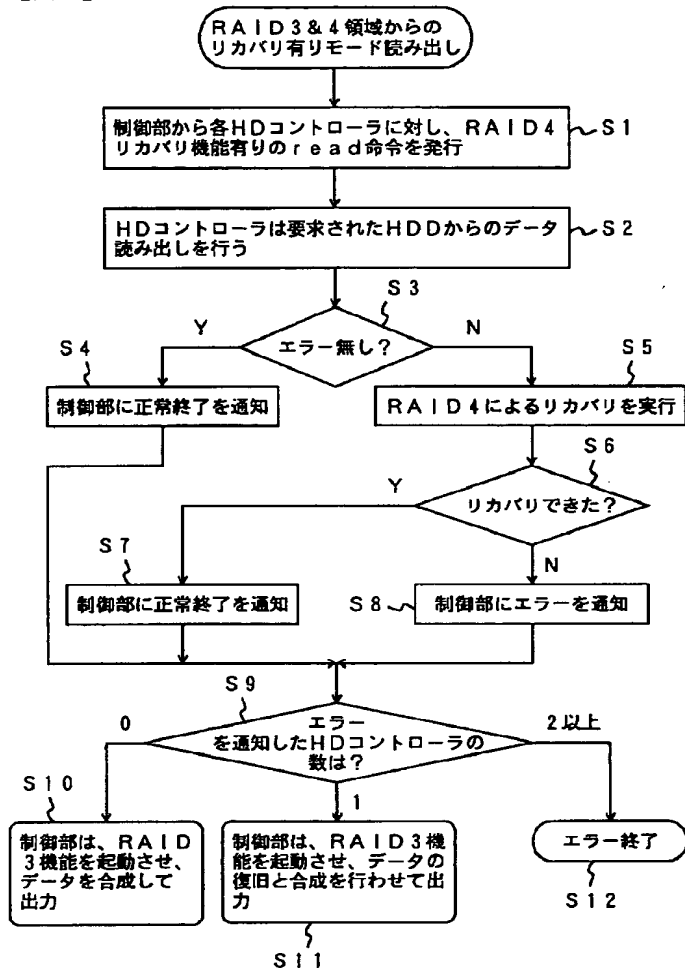
【図3】



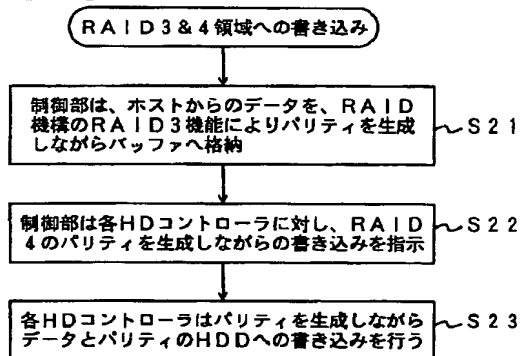
【図10】



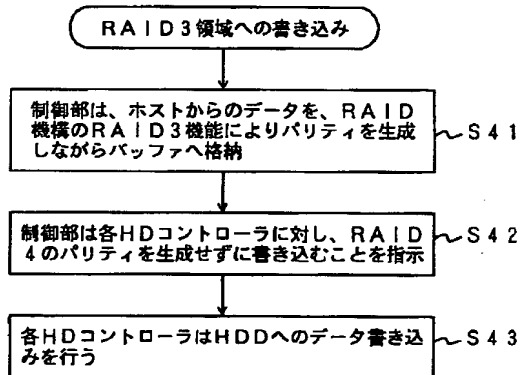
【図4】



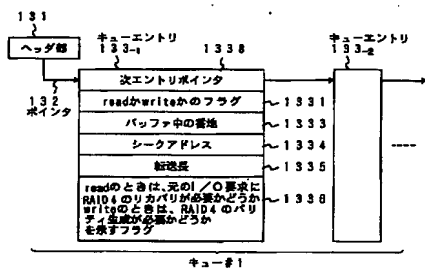
【図5】



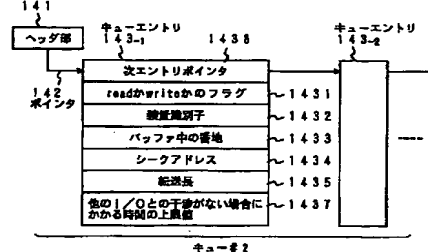
【図7】



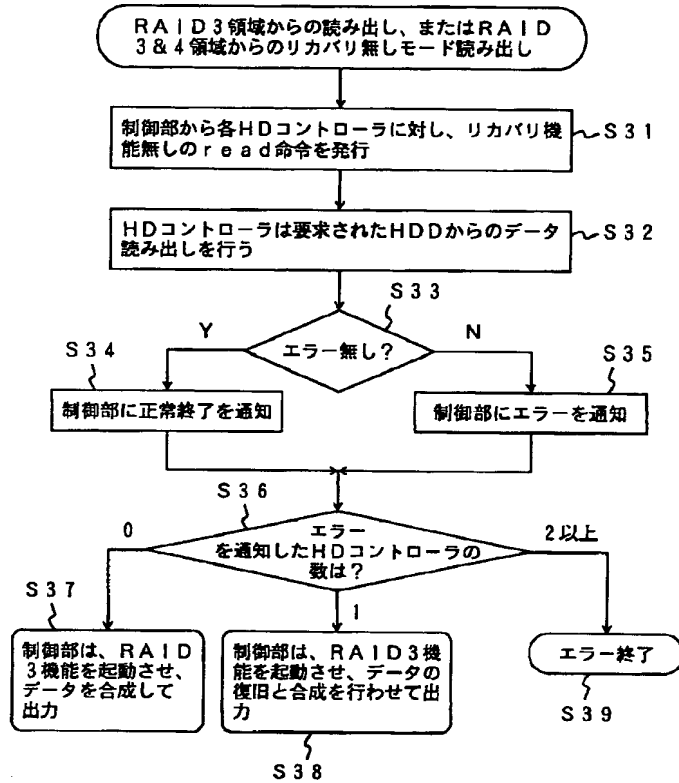
【図13】



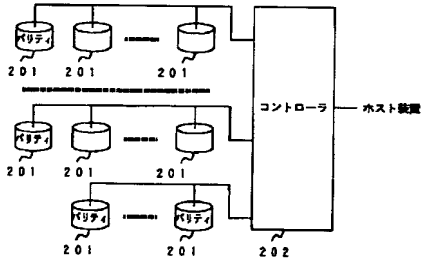
【図14】



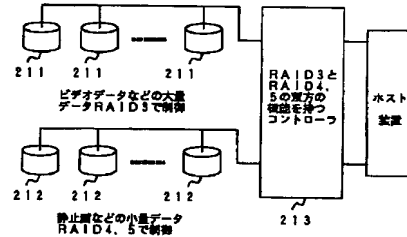
【図6】



【図20】

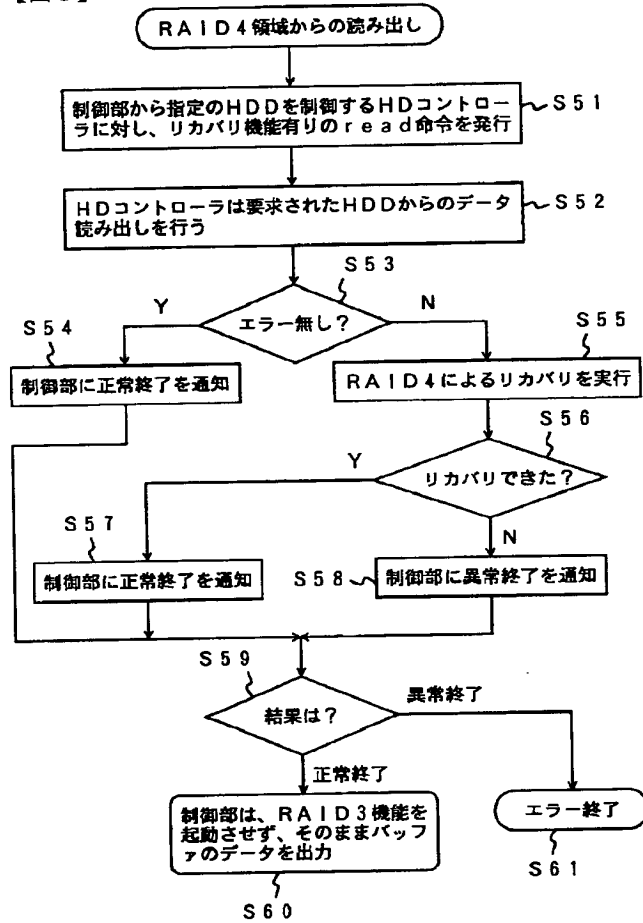


【図21】

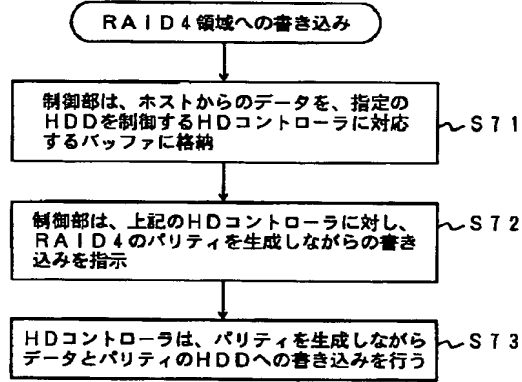




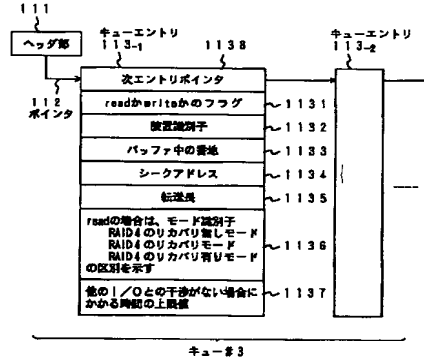
【図8】



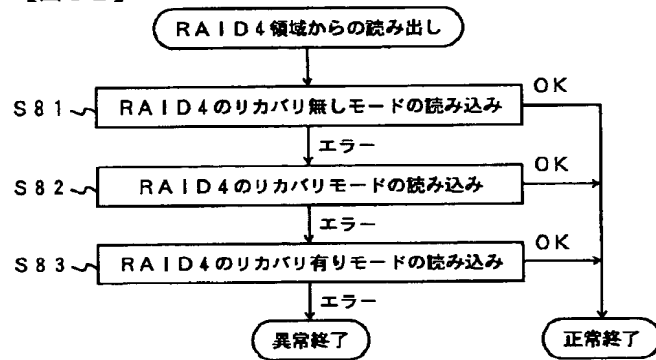
【図9】



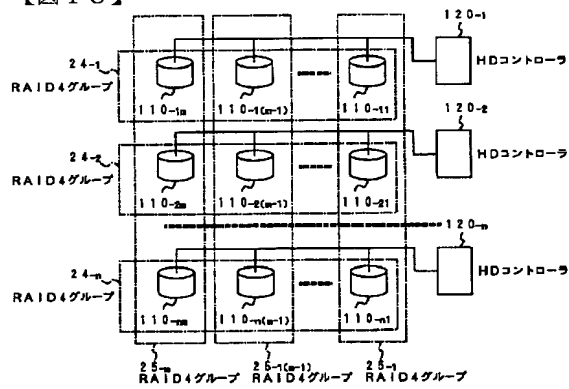
【図11】



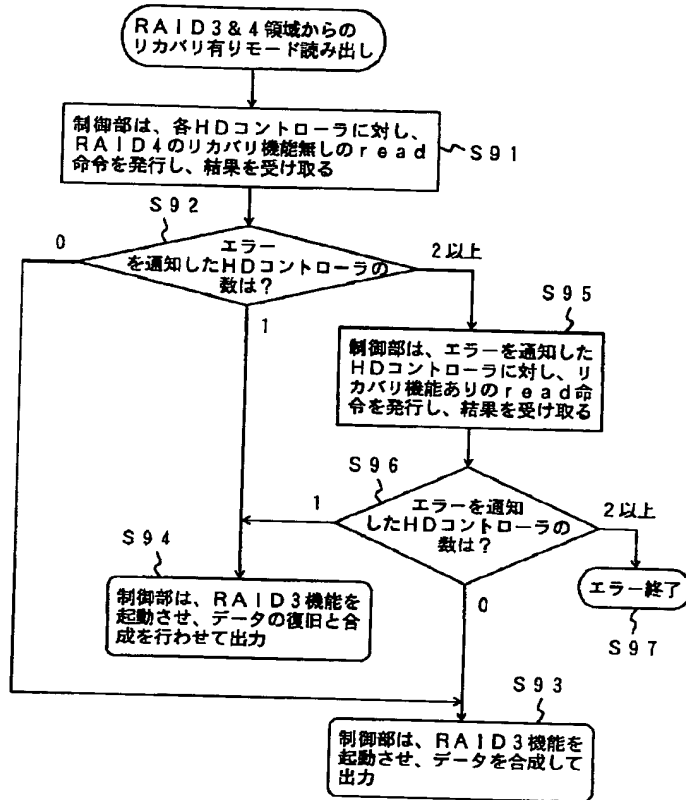
【図12】



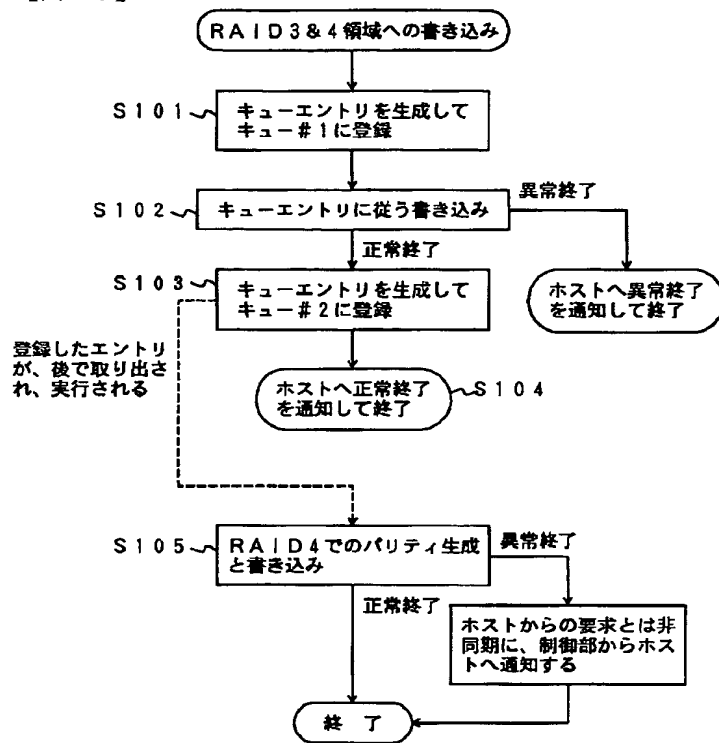
【図19】



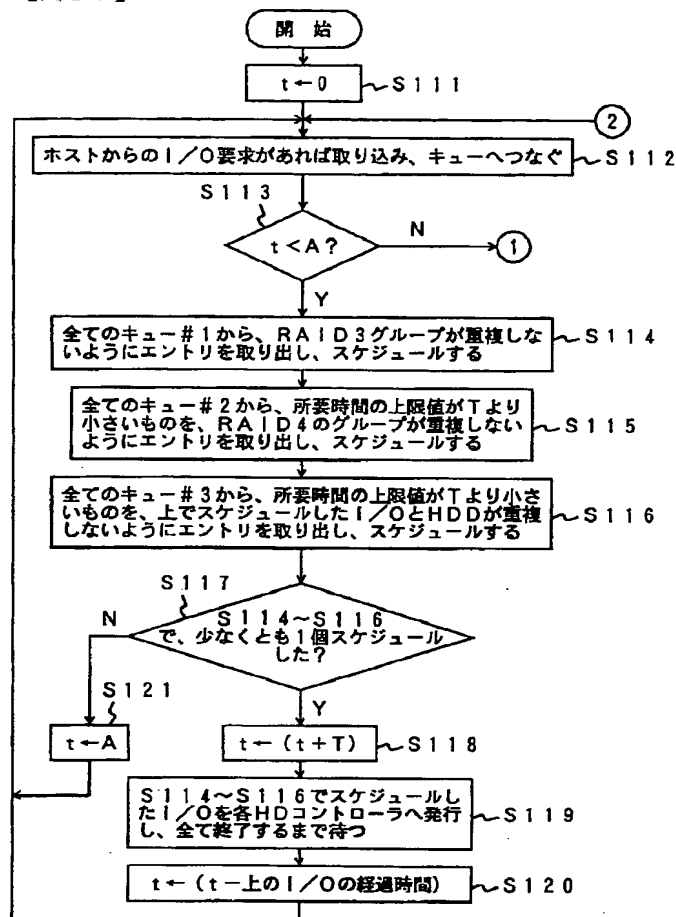
【図15】



【図16】



【図17】



【図18】

